



UNIVERSITY OF TRENTO

---

DOCTORAL SCHOOL IN PSYCHOLOGICAL SCIENCES AND  
EDUCATION  
(XXIII CYCLE)

PHD THESIS

# Uniqueness in Cognition

by

Barbara Bazzanella

ADVISOR:  
Prof. Paolo Bouquet  
University of Trento

SCHOOL COORDINATOR:  
Prof. Nicolao Bonini

December 2010



To my father

“To me, you are still nothing more than a little boy who is just like a hundred thousand other little boys. And I have no need of you. And you, on your part, have no need of me. To you, I am nothing more than a fox like a hundred thousand other foxes. But if you tame me, then we shall need each other. To me, you will be unique in all the world. To you, I shall be unique in all the world...”

The Little Prince (Chapter XXI)

## Abstract

A fundamental aspect of human cognition is that we construe the environment as including unique individuals that belong to various categories. An individual dog, for example, could simultaneously be a living being, a mammal or a poodle, but when it comes to things that are important to us - our dog Fido, our favorite restaurant, our spouse - we also represent the individuals themselves, not just the categories they belong to. Cognitive psychologists have made an extensive study of categories of objects but have had less to say concerning conceptions of individuals, i.e. *singular concepts*, and how they support our ability to uniquely identify individual entities in different situations.

The primary goal of this work is to investigate the nature and the functioning dynamics of *singular concepts* and explore how these concepts underlie *singular cognition*, i.e. the ability to identify a known entity, through perceptual or epistemic access to its memorial representation, and trace it as the same *unique* entity over time and change.

To perform such a process the cognitive system is confronted with a uniqueness problem. It needs to pick an individual entity out, secure a unique mental referential link with the entity and maintain that link over time and change.

We argue that singular concepts are the cognitive devices that are specialized for this function and we propose a model of singular cognition that has the notion of singular concept at its core. The main assumption of this model is that conceptual representations about individuals (i.e. singular concepts) represent a networks of unique files in memory which mediate the direct access to individual-specific knowledge and provide a unique mechanism of identification and reference for unique individuals. According to our model, the access to this system is not mediated by higher level representations (i.e. general concepts), neither is internally organized by these representations. On the contrary, it is subjected to its own functioning dynamics and it is organized through associative links which connect different individual concepts and causal links which maintain the conceptual history of an entity, by linking different states of the same singular concept, across time and change.

We can distinguish four main phases of our investigation about singular concepts which led to the proposed model of singular cognition.

- 1) In the first phase we investigated what is the preferential level of abstraction at which an individual entity is first identified (i.e. the entry point of recognition). Since any individual object can be identified at multiple levels of abstraction (e.g. a dog can be identified as a “dog”, more generally as “animal” or more specifically as “poodle” or “Fido”), the aim was to test the hypothesis that the singular concept of an object acts as the access node to the knowledge that the agent has about the object and this access is direct and not mediated by higher level concepts. Results from three experiments on visual recognition provided evidences in favor of this hypothesis, indicating that the entry level of identification of unique individuals is shifted to the most subordinate level of abstraction, i.e. the level of unique identity.

- 2) The second phase of this work explored how our semantic representations of individual things are accessed and how these representations are inter-linked with those of other individual things. This issue has been investigated through

a priming experiment which provided evidence in favor of a model in which singular concepts are organized by means of horizontal associative links instead of by vertical links with higher level representations.

3) In the third phase of our investigation we looked inside a singular concept and we explored which attributes people consider more relevant to uniquely identify entities belonging to different categories and determine the cognitive importance that individual attributes have in identifying these entities. We also explored which are the most relevant attributes that people use to identify entities in a specific task, i.e. the search for information about individual entities by means of keyword queries on the Web.

4) The last phase of the investigation concerned with the problem of how people judge the identity of entities over time and change. An experiment was conducted which explored how people evaluate the identity of entities over changes in their descriptions. The results of the study have been interpreted in the light of a causal model of the functioning of singular concepts in keeping the unique referential link with the entity across change.

Beyond the cognitive issues, this work is also motivated by the recent development of technological approaches to the problem of entity identification.

Since many identification problems which are addressed by a cognitive system have a counterpart in information systems which manage information about individual entities (e.g. to represent or extract information about unique individuals and manage individual-specific knowledge across time and change), the last goal of our work is to make an investigation of possible contributions that a cognitive study on the problem of individual identification can provide to technological applications. In particular we focused on the problem of entity identification in search systems. A model and an application for a specific technological problem, i.e. entity type disambiguation in Web-search queries, is described and its beneficial impact is evaluated.

In summary, the contribution of this work is twofold. On one hand, we provided new evidence on the nature of high-level cognitive mechanisms involved in entity representation and identification, suggesting new research issues on a field scarcely investigated in cognitive psychology. On the other hand, we provided concrete examples of how a better understanding of these processes at a cognitive level can improve the development of entity identification approaches in information systems, suggesting a middle ground where cognitive models and technological solutions can find the opportunity for integration, in particular in contexts characterized by interactions between humans and machines.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Mission Statement . . . . .	7
<b>2</b>	<b>Singular Concepts and Singular Cognition</b>	<b>9</b>
2.1	Singular Concepts: what they are and what they are not . . . . .	9
2.2	Functional Dynamics of Singular Concepts . . . . .	14
2.3	The problem of Singular Cognition . . . . .	17
<b>I</b>	<b>Background and State of the Art</b>	<b>19</b>
<b>3</b>	<b>Uniqueness in Cognitive Models</b>	<b>21</b>
3.1	Singular Representations and Reference in Psychology of Vision and Attention . . . . .	21
3.1.1	Object Files . . . . .	22
3.1.2	Visual Indexes . . . . .	25
3.1.3	Object Indexes and Object Concepts . . . . .	29
3.2	Singular Representations in Models of Face Recognition and Nam- ing . . . . .	31
3.3	Cognitive Theories of Object Identity . . . . .	41
3.3.1	Sortalist Approaches to the Problem of Singular Cognition	41
3.3.2	Non-sortalist Approaches to the Problem of Singular Cog- nition . . . . .	52
3.4	Proper Names as Index of Individual Identity . . . . .	57
<b>4</b>	<b>Neural Basis of Singular Concepts</b>	<b>63</b>
4.1	Neuropsychological Evidences . . . . .	64
4.2	Neuroimaging Evidences . . . . .	68
<b>5</b>	<b>The Problem of Identity in Information Systems</b>	<b>75</b>
5.1	Entity-level Information Integration . . . . .	76
5.2	Identity and Reference on the Semantic Web . . . . .	79
5.3	An Entity-Centric System for tracing the identity of entities on the Semantic Web . . . . .	81

<b>II</b>	<b>Novel Contributions</b>	<b>85</b>
<b>6</b>	<b>The Entry Point in the Identification of Individuals</b>	<b>87</b>
6.1	Introduction . . . . .	88
6.2	Objectives and Rationale of the Study . . . . .	92
6.3	Methodology . . . . .	95
6.4	Experiment 1: Entity Naming . . . . .	97
6.4.1	Method . . . . .	98
6.4.2	Results . . . . .	99
6.5	Experiment 2: Category-Verification Task . . . . .	104
6.5.1	Method . . . . .	105
6.5.2	Results . . . . .	107
6.6	Experiment 3: Identity Matching Task . . . . .	112
6.6.1	Method . . . . .	113
6.6.2	Results . . . . .	115
6.7	General Discussion . . . . .	117
<b>7</b>	<b>Associative and Semantic Priming in Recognition of Individuals</b>	<b>121</b>
7.1	Introduction . . . . .	121
7.2	Categorical and associative relatedness between entities and priming effects . . . . .	123
7.3	Objectives and Rationale of the Study . . . . .	131
7.4	An Entity Recognition Experiment . . . . .	133
7.4.1	Pilot Study: stimulus selection . . . . .	133
7.4.2	Method . . . . .	135
7.4.3	Results . . . . .	137
7.5	Discussion . . . . .	143
7.6	Implications for a Model of Entity Representation . . . . .	146
<b>8</b>	<b>Identification Relevance in Entity Representation</b>	<b>151</b>
8.1	Semantic Feature Norms Production for Individual Entities . . . . .	152
8.1.1	Method . . . . .	153
8.1.2	Results . . . . .	163
8.2	An Entity Search Experiment . . . . .	176
8.2.1	Method . . . . .	180
8.2.2	A Naive Bayes Model of Attribute Relevance . . . . .	182
8.2.3	Results . . . . .	183
8.2.4	Discussion . . . . .	196
<b>9</b>	<b>Tracing the Identity of Individual Entities</b>	<b>199</b>
9.1	Experiment 1: Mutability and Causal Distance Norms Production . . . . .	204
9.1.1	Method . . . . .	206
9.1.2	Results . . . . .	208
9.2	Experiment 2: Identity Decisions across Change . . . . .	209
9.2.1	Method . . . . .	211
9.2.2	Results . . . . .	213
9.2.3	Discussion . . . . .	219



<b>10 An Application for Entity Type Disambiguation in Queries using RDF Triples as Knowledge-Base</b>	<b>225</b>
10.1 Related Works . . . . .	228
10.2 The Entity Type Disambiguation Problem . . . . .	232
10.2.1 A simplified version of The Entity Type Disambiguation Problem . . . . .	233
10.3 A new approach for Entity Type Disambiguation . . . . .	235
10.4 PropLit: an application based on a index of RDF predicates . . .	238
10.5 Index Evaluation . . . . .	244
10.6 Conclusion . . . . .	251
<b>11 Conclusions and Future Work</b>	<b>255</b>
<b>A Experimental Materials used in the Entry Point Experiments</b>	<b>271</b>
A.1 Entry Point Experiment 2 . . . . .	271
A.2 Entry Point Experiment 3 . . . . .	273
<b>B Experimental Materials used in the Entity Recognition Experiment</b>	<b>275</b>
B.1 Entity Recognition Experiment . . . . .	275
<b>C Relevance Measures</b>	<b>277</b>
C.1 Feature Norms for Individual Entities . . . . .	277
C.2 Entity Search Experiment . . . . .	293
C.2.1 Attribute frequencies for the entity types of the entity search experiment . . . . .	293
C.2.2 Bayesian relevance measures for low-level entity types . .	298
C.2.3 Position Distribution of Attribute Types . . . . .	301
<b>D Mutability and Causality Ratings: Stimuli and Measures</b>	<b>304</b>
D.1 Entity profiles used to collect mutability and causality ratings . .	304
D.2 Mutability and Causality Ratings . . . . .	308
D.3 Entity Profiles used in Experiment 2 . . . . .	310
D.4 Response Distribution in Experiment 2 . . . . .	315
D.4.1 Causal Continuer Model Fit . . . . .	315
D.4.2 Naive Causal Model Fit . . . . .	325
<b>E PropLit Index</b>	<b>335</b>
E.1 Predicate-Entity Type Mapping used in the RDF index . . . . .	335
E.2 Top-50 RDF Predicates and their frequency . . . . .	337
<b>Bibliography</b>	<b>339</b>

# List of Figures

6.1	Percentage of basic level and subordinate level labels used in the naming task. . . . .	102
6.2	Category $\times$ Level of Categorization interaction. . . . .	103
6.3	Trial presentation sequence in the category verification task. . . .	106
6.4	Mean Reaction Times for the three categories of familiar entities, at the superordinate, basic and subordinate levels in the true condition. . . . .	108
6.5	Mean Reaction Times for the three categories of unfamiliar entities, at the superordinate, basic and subordinate levels in the true condition. . . . .	109
6.6	Mean Reaction Times for categorizing familiar and unfamiliar entities at superordinate, basic and subordinate levels in the TRUE condition. . . . .	110
6.7	Mean Reaction Times for categorizing familiar and unfamiliar entities at superordinate, basic and subordinate levels in the FALSE condition. . . . .	112
6.8	Trial presentation sequence in the identity matching task. . . . .	114
6.9	Results of the Experiment 3. . . . .	117
7.1	Trial presentation sequence in the entity recognition task (associative priming). . . . .	136
7.2	Trial presentation sequence in the entity recognition task (categorical priming). . . . .	137
7.3	Mean response times for the category Person Across by prime condition. . . . .	138
7.4	Mean response times for the Person Within category by prime condition. . . . .	139
7.5	Mean response times for the Artwork category by prime condition.	139
7.6	Mean response times for the Building category by prime condition.	140
7.7	Mean response time for the Product category by prime condition.	141
7.8	Mean response times for the Person Across and Person Within categories by prime condition. . . . .	141
7.9	Mean response times for the three object categories by prime condition. . . . .	142
7.10	Mean reaction times for face recognition and object recognition at the three levels of priming condition . . . . .	143
7.11	Associative and Categorical links between singular concepts . . .	147
8.1	Top-level categories and Ontological Commitments . . . . .	159

8.2	Self-evaluation of the participants regarding Internet and Semantic Web Experience . . . . .	163
8.3	Geographical provenance of participants of the entity search experiment. . . . .	180
8.4	Search interface used in the experiment to collect the participants'queries. In figure is shown the trial about a person search. . . . .	182
8.5	First five Google results for the query Q1=Silvio Berlusconi Mediaset. . . . .	190
8.6	First five Google results for the query Q2=Mediaset Silvio Berlusconi. . . . .	191
8.7	Probability distribution of attribute types for the first four positions in queries about Person. . . . .	192
8.8	Probability distribution of attribute types for the first four positions in queries about Organization. . . . .	192
9.1	Response frequencies of the response strategies used by participants (question 1) in the two experimental groups. . . . .	209
9.2	Response frequencies of the response strategies used by participants (question 2) in the two experimental groups. . . . .	210
9.3	Distribution of the causal distance differences between the continuers in the person tasks used in experiment 2. . . . .	212
9.4	Person tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model. . . . .	217
9.5	Person tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model. . . . .	218
10.1	Graphical representation of an RDF statement. . . . .	236
10.2	Snapshot of the list of ranked predicates returned by the basic search of the PropLit Index for the term "Barack". . . . .	241
10.3	Snapshot of the list of ranked predicates returned by the basic search of the PropLit Index for the term "San Salvador". . . . .	242
10.4	Snapshot of the output returned by the advanced search of the PropLit Index for the term "Barack". . . . .	242
10.5	Output of the basic (a) and advanced (b) modules for the query "Carlo Bonatti". . . . .	245
10.6	Output of the advanced search module for the query "New York USA". . . . .	246
10.7	Percentage of correct disambiguations (true positives) on single term queries. . . . .	247
C.1	Probability distribution of attribute types for the first four positions in queries about Person. . . . .	301
C.2	Probability distribution of attribute types for the first four positions in queries about Organization. . . . .	301
C.3	Probability distribution of attribute types for the first four positions in queries about Event. . . . .	302

C.4	Probability distribution of attribute types for the first four positions in queries about Artifact. . . . .	302
C.5	Probability distribution of attribute types for the first four positions in queries about Location. . . . .	303
D.1	Percentage of responses and fit of the Causal Continuer Model (person tasks 1-8). . . . .	315
D.2	Percentage of responses and fit of the Causal Continuer Model (person tasks 9-15). . . . .	316
D.3	Percentage of responses and fit of the Causal Continuer Model (organization tasks 1-8). . . . .	317
D.4	Percentage of responses and fit of the Causal Continuer Model (organization tasks 9-15). . . . .	318
D.5	Percentage of responses and fit of the Causal Continuer Model (event tasks 1-8). . . . .	319
D.6	Percentage of responses and fit of the Causal Continuer Model (event tasks 9-15). . . . .	320
D.7	Percentage of responses and fit of the Causal Continuer Model (artifact tasks 1-8). . . . .	321
D.8	Percentage of responses and fit of the Causal Continuer Model (artifact tasks 9-15). . . . .	322
D.9	Percentage of responses and fit of the Causal Continuer Model (location tasks 1-8). . . . .	323
D.10	Percentage of responses and fit of the Causal Continuer Model (location tasks 9-15). . . . .	324
D.11	Percentage of responses and fit of the Naive Causal Model (person tasks 1-8). . . . .	325
D.12	Percentage of responses and fit of the Naive Causal Model (person tasks 9-15). . . . .	326
D.13	Percentage of responses and fit of the Naive Causal Model (organization tasks 1-8). . . . .	327
D.14	Percentage of responses and fit of the Naive Causal Model (organization tasks 9-15). . . . .	328
D.15	Percentage of responses and fit of the Naive Causal Model (event tasks 1-8). . . . .	329
D.16	Percentage of responses and fit of the Naive Causal Model (event tasks 9-15). . . . .	330
D.17	Percentage of responses and fit of the Naive Causal Model (artifact tasks 1-8). . . . .	331
D.18	Percentage of responses and fit of the Naive Causal Model (artifact tasks 9-15). . . . .	332
D.19	Percentage of responses and fit of the Naive Causal Model (location tasks 1-8). . . . .	333
D.20	Percentage of responses and fit of the Naive Causal Model (location tasks 9-15). . . . .	334

# List of Tables

6.1	List of familiar artifacts used in Experiment 1 . . . . .	98
6.2	List of unfamiliar artifacts used in Experiment 1 . . . . .	99
6.3	Percentage Frequencies by Object category and Level of abstraction	101
6.4	Percentage Frequencies by Object category and Level of abstraction: aggregated responses . . . . .	101
6.5	Percentage of correct TRUE (CT) and correct FALSE (CF) responses by category. . . . .	107
6.6	Mean Reaction Times for the TRUE responses as a function of Category (familiar vs. unfamiliar) and Category Level (superordinate, basic and subordinate). . . . .	108
6.7	Mean Reaction Times for the FALSE responses as a function of Category (familiar vs. unfamiliar) and Category Level (superordinate, basic and subordinate). . . . .	111
6.8	Mean RTs in milliseconds (and standard errors of the mean) by Object Category and Prime Type . . . . .	115
7.1	Mean Reaction Times (RT) in milliseconds (and Standard Errors (SE)) and Accuracies (AC) for Conditions of the Entity Recognition Experiment . . . . .	137
8.1	Categories and Subcategories used in the experiment. . . . .	161
8.2	Geographical provenance of participants. . . . .	162
8.3	Normalization examples. 1=structural normalization; 2= morphological normalization; 3=semantic normalization . . . . .	164
8.4	Features and production frequencies for the category <i>Person</i> . . . . .	165
8.5	Features and production frequencies for category <i>Person</i> : aggregated data (i.e. English and Italian). . . . .	166
8.6	Local Dominance for selected top-level categories. We marked with an * the attributes that appear in the first 5 positions of dominance also in the analysis which considered only the neutral categories. . . . .	171
8.7	Entity types and subtypes used in the entity search experiment. . . . .	182
8.8	Two-step Preprocessing . . . . .	183
8.9	Bayesian Relevance: top-level entity types . . . . .	186
8.10	Attribute Types: examples . . . . .	186
8.11	Bayesian Relevance: PERSON . . . . .	187
8.12	Confusion matrix: learning set. . . . .	188
8.13	Learning set evaluation. TP=true positive; FP=false positive . . . . .	189
8.14	Test set evaluation . . . . .	189

8.15	Test-set Evaluation of the NBM. . . . .	193
8.16	Test-set Evaluation of the EBM . . . . .	193
8.17	Performance Measures . . . . .	195
8.18	Relative performance . . . . .	196
9.1	Causal Continuer Model Fit . . . . .	215
9.2	Naive Causal Model Fit . . . . .	219
10.1	Entity Type Disambiguation example . . . . .	247
10.2	Performance of the advanced search module on single term queries.	248
10.3	Detection and Disambiguation Effectiveness . . . . .	249
10.4	Baseline performance . . . . .	250
A.1	Stimuli and categories words used in experiment 2 . . . . .	272
A.2	Stimuli and categories words used in experiment 3 . . . . .	274
B.1	Prime words and Stimuli used in the entity recognition experiment	276
C.1	Features and production frequencies for PERSON . . . . .	278
C.2	Features and production frequencies for PERSON: aggregated data (i.e. English and Italian). . . . .	279
C.3	Features and production frequencies for ORGANIZATION . . . . .	280
C.4	Features and production frequencies for ORGANIZATION: aggre- gated data (i.e. English and Italian) . . . . .	281
C.5	Features and production frequencies for EVENT . . . . .	282
C.6	Features and production frequencies for EVENT: aggregated data (i.e. English and Italian) . . . . .	283
C.7	Features and production frequencies for ARTIFACT . . . . .	284
C.8	Features and production frequencies for ARTIFACT: aggregated data (i.e. English and Italian) . . . . .	285
C.9	Features and production frequencies for LOCATION . . . . .	286
C.10	Features and production frequencies for LOCATION: aggregated data (i.e. English and Italian) . . . . .	287
C.11	Relevance Measure for PERSON . . . . .	288
C.12	Relevance for ORGANIZATION . . . . .	289
C.13	Relevance for EVENT . . . . .	290
C.14	Relevance for ARTIFACT . . . . .	291
C.15	Relevance for LOCATION . . . . .	292
C.16	Attribute frequencies in PERSON queries. . . . .	293
C.17	Attribute frequencies in ORGANIZATION queries. . . . .	294
C.18	Attribute frequencies in EVENT queries. . . . .	295
C.19	Attribute frequencies in ARTIFACT queries. . . . .	296
C.20	Attribute frequencies in LOCATION queries. . . . .	297
C.21	Bayesian Relevance: PERSON . . . . .	298
C.22	Bayesian Relevance: ORGANIZATION . . . . .	298
C.23	Bayesian Relevance: EVENT . . . . .	299
C.24	Bayesian Relevance: ARTIFACT . . . . .	299
C.25	Bayesian Relevance: LOCATION . . . . .	300
D.1	Mean (and SD= standard deviation) Mutability and Causality Ratings . . . . .	310

D.2	Person Profiles used in the experiment2 . . . . .	311
D.3	Organization Profiles used in the experiment2 . . . . .	312
D.4	Event Profiles used in the experiment2 . . . . .	313
D.5	Artifact Profiles used in the experiment2 . . . . .	314
D.6	Location Profiles used in the experiment2 . . . . .	315
E.1	Predicate-Entity Type mapping schema . . . . .	337
E.2	Top-50 RDF Predicates and their frequency in the 1 billion triple store. . . . .	338





# Acknowledgements

This thesis is about unique individuals but it also the product of the support and collaborative input of many “unique” people. It is my pleasure to have the opportunity to express my gratitude to many of them here.

First and foremost I want to thank my PhD supervisor prof. Paolo Bouquet for his trust and support. I would like to thank him for advising me during the course of this work, but also for giving me the freedom to find my own way. His guidance and truly scientist intuition helped me in all the time of research and writing of this thesis and contributed significantly to my growth as a student, as a researcher and as a person.

Special gratitude goes to prof. Lance Rips for the opportunity to carry out part of this project in his Higher-Level Cognition Lab at the Northwestern University of Chicago (USA). He was always open for precious suggestions and scientific discussions and accepted me as a full member of his group. In particular, many thanks go to Winston Chang. My work at the Northwestern University would not have been possible without his assistance and support, as well as his friendship and good advice.

I would like also to express my sincere gratitude to Dr. Giovanni Tummarello for giving me the opportunity to work in his group at the Digital Enterprise Research Institute (Galway, Ireland). The group has been a source of friendships as well as good advice and collaboration. In this context, a very special thanks to Michele Catasta for his help and support during the realization of the PropLit index.

I would also like to thank my colleagues in the Okkam Lab, particularly Heiko Stoermer, Angela Fogarolli, Stefano Bortoli, Massimiliano Vignolo, Daniel Giacomuzzi, Sven Buschbeck, Xin Liu and George Giannakopoulos for our debates, exchanges of knowledge, skills, and venting of frustration during my PhD program, which helped enrich the experience. Thanks all of you for the great time I had in our group, as well as for the many coffees, sandwiches, teas, and funny conversations we’ve had over the last years.

For this work, data were essential. I collected a lot of data. Many people helped with this, for which I would like to thank them wholeheartedly. Hundreds of people participated to my experiments. Without their generosity there would be nothing to work with.

I also thank my friends (too many to list here but you know who you are!) for their understanding, endless patience and encouragement during these years. Among them, I would like to thank a very special person, Matteo, because he taught me a very important thing in life: we cannot change the cards we are dealt, just how we play the hand.

Finally, I want to express my deepest love and gratitude to the most impor-

tant people in my life, my parents and my family, who have fully supported me in all imaginable ways. I would never have achieved what I have had without your continual understanding, support and encouragement. Thank you: Mom, Dad, Luca, Manuela, Luciano, Silvia, Leonardo, Aronne, Mirko, Sonia, Chiara, Noemi, Irene, Silvana and Camilla. Thank you for never losing trust in me.

# Chapter 1

## Introduction

A fundamental aspect of human cognition is that we conceptualize the reality as including unique entities - such as specific places, persons, objects - that belong to various categories. An individual dog, for example, could simultaneously be a living being, a mammal or a poodle, but when we identify a specific dog as our dog “Fido” we access the mental representation of the individual itself not just the categories it belongs to. The processes involved in identifying an object at these two levels of abstraction (as a member of a general category or as a unique individual) are indeed quite different, as they are, we assume, their underlying memory representations. When classifying an object as a member of a category, we need to ignore the very information that is required to distinguish individual exemplars of the category and we need to connect the object with a general conceptual representation which cluster features largely shared by the members of the category. On the contrary, when we identify an object as a unique individual we activate unique semantic associations that are distinctive of that particular object compared to the other category members. We refer to the cluster of unique semantic associations linked to an entity as a *singular concept*, while we name *singular cognition* the complex of cognitive processes that allow a cognitive agent to identify a known entity, through perceptual or epistemic access to its memorial representation, and trace it as the same unique entity perceived or known at successive moments in time. To perform such a process the cognitive system is confronted with a uniqueness problem. It needs to pick an individual entity out, secure a unique mental referential link with the entity and maintain that link over time and change. We argue that singular concepts are the cognitive devices that are specialized for this function, providing a unique referential link between the entity and its memorial representation.

For years, cognitive psychologists have made an extensive study of categories of objects and their mental representations (i.e. general concepts), but have

put less effort studying how people represent unique individuals and how these representations (i.e. singular concepts) support our ability to identify these entities in different situations. Nonetheless, the identification process is crucial to daily life. We need to correctly recognize and identify all the individual entities relevant to our own existence (people, pets, places, objects and so on) and successfully track those individuals over time and change. When these abilities are compromised, the consequences may be devastating and a complex array of neuropsychological deficits have been documented at various stages of the identification process (a stirring example is described by Oliver Sacks in his famous book “The Man Who Mistook His Wife for a Hat” [205]).

The identification process has often been treated in a perceptual context (e.g. face recognition), since perceptual factors play a fundamental role in identifying objects. For example, I can recognize a person as familiar at a crowded party by her perceptual appearance (e.g. her face, voice, clothing) and then fully identify the person by retrieving semantic information about her, including her name.

However, there are contexts in which perceptual information is scarce or insufficient to ensure correct identifications. If I see the same friend after a long time, the stock of perceptual informations which are part of my representation of her may be too dated to support the identification process and the full identification can be obtained only by acquiring extra information, e.g. during the conversation. In other cases perceptual information can be completely absent. For example, while reading a news item about a traffic accident, I can suspect that the person involved in the accident is a classmate whom I lost touch with a long time ago. In this situation, I have to decide about the identity of this individual using only the information reported in the article. In these cases, we can only rely on conceptual histories and higher level knowledge about individuals must come into play to allow the identification process.

We argue that this knowledge is stored in specialized mental files, which we refer to as *singular concepts*, which bind together our information about the individuals they are about and individuate our cognitive perspective on those individuals. However, singular concepts are not only vehicles for storing information about a particular individual, but they serve as mental identifiers which create the unique referential link between an object in the world and its mental representation in the cognitive system.

In this sense, singular concepts represent the core of the identification process both in perceptual contexts and in contexts in which perceptual information is scarce or not available at all. Identification depends, indeed, on a variety of cognitive means for information acquisition, such as perception, reasoning, communication and so on. The acknowledgment of this variety of means requires distinguishing two ways to access singular concepts: *perceptual* or *bottom-up* and

*epistemic* or top-down.

A mental file can be accessed via a *bottom-up* way by a perceptual stimulus. In this case the individual is present in a sensory field of the agent’s perceptual systems and the perceptual input activates the corresponding mental representation, through a direct match with the perceptual information stored in the concept. Alternatively, a mental file can be accessed via a *top-down* way in cases in which the target individual cannot be perceived, but can be identified on the basis of indirect information gathered by such sources as memory, reasoning or communication.

Even though many cognitive models of human knowledge assume the existence of mental representations of unique individuals <sup>1</sup>, the nature of these representations has been less investigated or has been considered less relevant to understand the identification process compared to higher level conceptual representations (as assumed, for example, by sortalist approaches to identity [268]).

The first aim of the present work is to provide a contribution to fill this gap, by proposing a general model of singular cognition based on a system of mental unique representations which ensure the agent’s individuation of unique objects through different contexts, time and change. In other words the focus is on the mechanisms and the cognitive devices of singular cognition.

We can distinguish four main phases of our investigation about singular concepts. We start with the study of how an individual entity first makes contact with its underlying memorial representation through a bottom-up way (i.e. from a perceptual stimulus to a singular concept); then, we investigate how singular concepts are organized and interrelated each other in the conceptual system; we then pass to explore some aspects of the internal organization of the knowledge stored in a singular concept and we consider the top-down access to singular concepts in a specific identification task; we finally investigate how singular concepts underlie the mechanism of tracing entities through time and change. More precisely:

1. The first phase of our research aims to investigate what is the preferential level of abstraction at which an individual entity is first identified (i.e. the

---

<sup>1</sup>Unique representations of individuals are assumed in memory studies [3, 195]; neuropsychology studies have suggested specialized neural mechanisms devoted to evoking memories about unique members of categories [54] and specific impairments for unique entity-information are reported in literature [92, 90]. The existence of representations of unique individuals distinct from those of general categories is also expressed in conceptual semantics in the distinction between “tokens” and “types” [112] and exemplar models of categorization [159] make the same distinction, proposing that people represent categories by means of representations of unique individuals. Finally, the most accepted cognitive models of face recognition and naming [35, 38] assume the existence of identity nodes in memory which store semantic knowledge about individuals and are accessed to fully identify known persons.

entry point). In particular, since any individual object can be identified at multiple levels of abstraction (e.g. a dog can be identified as a “dog”, more generally as “animal” or more specifically as “poodle” or “Fido”), the aim is to test the hypothesis that the singular concept of an object acts as the access node to the knowledge that the agent has about the object and this access is direct and not mediated by higher level concepts.

2. The second phase of this work explores how our semantic representations of individual things are organized and accessed and how these representations are inter-linked with those of other individual things. In particular we test two alternative views about the organization of singular concepts in semantic memory. A *categorical view* which holds that memory representations of unique entities are interconnected by belonging to common categories and an *associative view* which holds that relationships between entities can be represented by networks of associative links but not by membership of a common category.
3. In the third phase of our investigation we look inside a singular concept and we explore which attributes people consider more relevant to uniquely identify entities belonging to different categories and determine the cognitive importance, or weight, that individual attributes have in identifying these entities. We also explore which are the most relevant attributes that people use to identify entities in a specific task, i.e. the search for information about individual entities by means of keyword queries on the Web. In this phase we explore the hypothesis that information within a singular concept is organized in terms of identification relevance and that the notion of relevance is, at least in part, contextual dependent.
4. Finally, the last phase of the research concerns with the problem of how people judge the identity of entities over time and change. Because individuals can change some of their properties while persisting as the same individuals, the singular cognition system needs a function of tracking a changing entity by performing specialized updating operations, which maintain the referential link with its singular concept. In this phase of our research we explore how causal mechanisms come into play to connect the possible states of a singular concept at successive moments in time.

On the basis of the results of the four phases described above, we finally propose a cognitive model for singular cognition that has the notion of singular concept at its core. The main assumption of this model is that conceptual representations about individuals (i.e. singular concepts) represent a networks of unique files in memory which mediate the direct access to individual-specific knowledge

and provide a unique mechanism of identification and reference through time. According to our model, the access to this system is not mediated by higher level representations (i.e. general concepts), neither is internally organized by these representations, but it is subjected to own functioning dynamics based on associative links which connect different individual concepts and causal links which connect different states of the same singular concept across time and change.

Beyond the cognitive issues, this work is also motivated by the recent development of technological approaches to the problem of entity identification.

Much of information in current decentralized network-based systems - including the Web and its evolving extension, the Semantic Web <sup>2</sup> - is about individual entities and recently we are assisting to the transition from the centrality of documents to that of entities as atomic objects of information. This transition has been recently marked in research on knowledge representation and integration by the passage from approaches more focused on high level representations - i.e. ontologies with special focus on the T-Box part of the ontology, which defines concepts and its relations - to the emergence of entity-centric approaches which focus on the instances which populate ontologies, realizing what has been recently called the “A-Box revenge” [235]. However it is worth to note that this is just a recent phenomenon and for years research on knowledge representation and integration was research on general categories and their relations, as well as nearly all the research on concepts in cognitive psychology was research on general categories of objects.

Contrary to the traditional trend, today, a big effort is made to allow the information integration across multiple heterogeneous sources and the idea that identifying entities is at the core of this effort is increasingly diffuse, representing also one of the main pillars of the Semantic Web.

Ideally, the information integration could be obtained by uniquely identifying entities in all the local nodes of a distributed system. However, the solution proposed by the Semantic Web to extend the use of a URI (Uniform Resource Identifier) to identify not just web pages, but any resource on the Web [20, 19], does not ensure that the same entity is consistently assigned the same URI across different sources.

Several theoretical and technological solutions to the problem of identification in the Semantic Web - referred as “identity crisis” by [44] - have been

---

<sup>2</sup>The Semantic Web vision was conceived by Tim Berners-Lee, the inventor of the World Wide Web, as an extension of the current Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [19]. The majority of today’s World Wide Web’s content is designed for humans to read and understand, not for machines and computer programs to manipulate meaningfully. Machines have no reliable way to process the semantics of the Web documents. The Semantic Web will bring structure to the meaningful content of Web pages through the use of standards, markup languages and related processing tools with the intent to facilitate information integration, reuse and exchange, across application and systems.

recently proposed [101, 26].

A possible solution to the entity identification problem can be found within the OKKAM project [28, 29]. The goal of OKKAM is to develop an Entity Name System (ENS) for the (Semantic) Web, a web-scale infrastructure which can make sure that the same web entity is referred to through the same URI across any type of content, format, application. The ENS has a repository for storing entity identifiers along with some small amount of descriptive information for each entity. When a request for an entity is submitted, the ENS decides if a URI for this entity is already available in the repository; if it is, then the ENS will return its URI, otherwise it will issue a new URI which will be stored in the repository.

The development of an ENS leads inevitably to issues of entity representation and identification that are common to any systems that represent and manage information about entities.

The problem to define what counts as an entity is a fundamental issue. Which are the atomic objects that needs to be referenced and distinguished from other objects in an information system?

To identify something, it is necessary to distinguish it from other things, which leads to the question how an entity is supposed to be described in a way that sufficiently distinguishes it from all the other entities. Which is the most important information that allows to identify an entity?

Identity decisions, i.e. the decision if two entity descriptions refer to the same entity, are performed mainly on the information about the entity stored in the system. However the information about an entity can change across time. Which is the information more likely to change over time? What is the influence of entity change on identity?

These questions show that many cognitive issues mentioned above are relevant also from a technological perspective.

There is a strong parallelism between the identification needs in a cognitive system and those in a entity-centric system. Moreover many of the dynamics which govern the functioning of singular concept have a counterpart in the functioning of entity-centric systems. In both cases, the system stores, accumulates and updates information about individual entities creating singular representation of them. A special kind of referential mechanism between singular representations in the system and the corresponding token elements in the world is required to recognize an entity as a familiar one, to access to its entity specific information (in a bottom-up or top-down way) and to fully identify the entity across successive moments in time.

Another important aspect that connects the identification issues of a cognitive agent with those of information systems is that in their interactions with



these systems people are more and more faced with identifying and searching information about individual entities. Therefore, the success of these interactions strongly depends on the ability for the automatic system to correctly interpret the singular cognition act, i.e. the identification act, of the user and return the information about the intended target.

Therefore, the second aim of the present work is to provide evidence of how a cognitive study on the identification problem can contribute to answer analogous questions in a technological context and inspire possible solutions to some of the most crucial issues about entity identification in entity-based systems, such as entity representation or entity resolution <sup>3</sup>.

5. The fifth and last phase of the present work aims to map some of the main cognitive issues about entity identification into corresponding technological issues, with the aim to show how the solutions adopted by the cognitive system can inspire and improve models and algorithms for identification which can be adopted in information systems. A practical example of this is reported in Chapter 10 where we describe a model and an application for entity type disambiguation of keywords in entity search.

## 1.1 Mission Statement

In this work we argue that a model of how individual entities are analyzed and represented by the cognitive system for the identification process will have to provide a system that does more than construct a conceptual representation of these entities. Such a model, which we might call a model of singular cognition, will also have to provide a special kind of referential connection between the elements of the mental representation and certain token elements in the world, a connection that is unmediated by higher-level conceptual representations, i.e. general concepts. We argue that this connection is secured by special mental representations, which we refer to as “*singular concepts*”, that provide a system of unique mental identifiers for unique entities.

The main goal of this work is to explore the nature of this system and propose a model of how people identify individual entities and trace the identity of these entities through time and change. To this end, we will address the following objectives:

---

<sup>3</sup>Entity Resolution (ER) is an important information integration problem: The same “real-world entities” are referred to in different ways in multiple data records. For instance, two records on the same person may provide different name spellings, and addresses may differ. The goal of ER is to “resolve” entities, by identifying the records that represent the same entity and reconciling them to obtain one record per entity.

1. to explore the first point of access to entity-specific information, stored in singular concepts, during the (visual) recognition of unique entities;
2. to investigate how semantic representations of individual entities are organized and how these representations are inter-linked with those of other individual entities;
3. to investigate the internal structure of singular concepts, showing how semantic features represented within a singular concept may have different importance in concept representation and provide evidence of which features people consider most important to uniquely identify individual entities in different tasks;
4. to study how causal factors are involved in shaping concepts of individuals and explore how people make use of causal information to identify objects across time and change.
5. to explore the parallelism between identification problem in a cognitive system and the same problem in an information system and provide evidence for possible applications and benefits in developing methods for automatic entity identity management.

In summary, the contribution of our work is twofold. On one hand, we aim to examine the nature of high-level cognitive mechanisms involved in entity representation and identification, revealing new research issues on this topic in cognitive psychology. On the other hand, we aim to explore how a better understanding of these processes at a cognitive level, can improve the development of entity identification mechanisms required by systems that manage automatically the identity of represented entities.

## Chapter 2

# Singular Concepts and Singular Cognition

Many models of human cognition assume the existence of mental representations of unique individuals<sup>1</sup>. However, the nature of these representations and their functioning dynamics have been poorly investigated in cognitive psychology, in particular with reference to the identification mechanisms involved in singular cognition. In this chapter we explain what is a singular concept (and what is not a singular concept), which are its properties, how it functions and why we need such a representation mechanism which identify and keeps track of individual objects in the world across time.

### 2.1 Singular Concepts: what they are and what they are not

Our knowledge of the world is mediated by two kinds of activities: 1) perceptual activities, providing us with information about the external world; and 2) conceptual activities, allowing us to have internal representations of various categories of objects.

---

<sup>1</sup>For example, many models of semantic memory represent individuals differently than classes [3, 195] and neuropsychology studies have suggested specialized neural mechanisms devoted to evoking memories about unique members of categories. [54]. The existence of representations of unique individuals distinct from those of general categories is also expressed in conceptual semantics in the distinction between “tokens” and “types” [112] and exemplar models of categorization [159] make the same distinction, proposing that people represent categories by means of representations of unique individuals. Finally, the most accepted cognitive models of face recognition and naming [35, 38] assume the existence of identity nodes in memory which store semantic knowledge about individuals and are accessed to fully identify known persons.

Nearly all research on conceptual activities in cognitive psychology is research on categories of objects, e.g. categories as “dog”, “chair” or “car”. But when it comes to things that are important to us - people, works of art, buildings, places - we understand that much of human knowledge is about “individuals” and we have representations of these individuals, not just of the categories they belong to. This lead us to distinguish between two kinds of conceptual representations: *singular concepts* and *general concepts*.

We use the term *singular concept* to denote a cognitive representation of a unique individual, and we contrast singular concepts with *general concepts*, which are representations of categories. A representation of The Leaning Tower of Pisa is a singular concept in these terms, but our representation of (the category of) buildings or towers is a general one. There are many important differences between general and singular concepts and these differences explain some of the most peculiar properties of singular concepts.

- First of all, every (individual) entity may be identified by more than a single general concept (e.g. my dog can be identified as an “animal”, as a “dog” or as a “poodle”) and several entities can be identified by the same general concept, even though some entities can be better exemplars than others [196], e.g. my dog and the dog of my neighbor are both identified as “dogs”. On the contrary, a unique singular concept is build in memory which represents a specific individual and an entity can be fully identified as that unique entity (e.g. Fido), only by activating the corresponding singular concept <sup>2</sup>. This means that there is a one-to-one relationship between the individual in the world and its representation in memory. We refer this property of singular concepts as *conceptual uniqueness*.
- Many cognitive theories assume that concepts can be considered as organized structures of semantic features [197, 226, 157] and models have been proposed to capture the relative importance of different semantic features to the meaning of a concept [207]. We argue that also singular concepts can be modeled adopting a feature-based approach. However, an important difference between general and singular concepts lies in the kinds of features which represent the core meaning of a concept, i.e. the most useful features in discriminating the concept from those similar to it. In case of a general concept the most relevant attributes are those highly shared by the members of the category, whereas in case of a singular concept the most relevant attributes are those highly distinctive of that particular individual. This means that when identifying an object by mens of a general concept we focus on properties shared by other members of the same

---

<sup>2</sup>In the course of this document we will discuss possible exceptions to this property.

category, ignoring the very information that is required to distinguish the object from the other exemplars of the same category. On the contrary, when identifying an object by means of a singular concept we need to discard the information highly shared by the exemplars of the category and focus on the distinctive information of that particular object. We name this property *conceptual distinctiveness*.

- A third important aspect deals with the following question: why do we have singular representations for certain particulars and not for others? or in other words, why do we represent certain entities only as members of general categories, but we represent other entities as unique members of singular concepts? This question reminds a similar issue about proper names that - as noted by Robin Jeshion in her paper on the significance of proper names [115] - was addressed (among other philosophers) by John Locke that discussed the issue in his *Essay Concerning Human Understanding* [139]. Locke wondered “why isn’t it the case that all things have proper names?” The Locke’s answer was that it is “psychologically” impossible for a human “to frame and retain distinct ideas of all the particular things we meet with: every bird and beast men saw; every tree and plant that affected the senses, could not find a place in the most capacious understanding” [139, book 4, ch 3, §2 ]. The same constrain can be applied to singular concepts. Since the cognitive system has limited resources, it would be cognitively impossible to manage singular concepts for all the unique entities which a cognitive agent can meet with.

Jeshion noted that the second answer suggested by Locke was that, “even if it were possible to name every particular, it would be useless for communication have a proper name for each of them” [115], because naming particulars never communicated about with others could not serve the end of communication, or to use the words of Locke “it would not serve to the chief end of language” . This second answer suggests a second reason of why we have singular representations for certain particulars and not for others. We represent unique entities by means of singular concepts if this is useful for a certain function. While the function of proper names is to fix the reference in communication, functioning as longterm, interpersonally available linguistic labels of their referents, singular concepts are conceptual devices to uniquely identify individuals within the conceptual system of a human agent. Therefore, we argue that singular concepts are initiated for those entities we need to mark as “unique” in semantic memory. There are at least two reasons why we need to mark this uniqueness.

First, we need to single out an individual from all other similar individuals.

I need to distinguish my dog Fido from all the other golden retriever which I meet with. However, it is unlikely that I need to single out the dog that I saw playing in the park last evening from all the other dogs, unless there is a particular reason (e.g. I was bitten by him).

Second, as argued by Jeshion in the case of proper names [115], we need to mark the significance of an individual. According to the “relevance view” proposed by the author “proper names and their associated mental representations are” non only devices of direct reference but, “additionally, and by their nature, markers of their referents’ significance”. In the same way, it is likely that we possess singular representations for certain “special” particulars because we introduce singular representations especially for those particulars we regard as having intrinsic value, beyond their value as an instance of a certain category, and we do so because we wish to signal that “uniqueness”. We note, however, that we can signal the uniqueness and the relevance of an individual by building a singular concept of it , without marking the concept with a proper name label. There are objects that are relevant for us but we do not refer to them by using proper names. My Iphone, for instance, is not just a member of the mobile phone category. I have a very distinct representation of it in my semantic memory that contains, for example, information about the nice applications that I installed on it, the color of the bumper, the picture that I’m using as wallpaper and so on. In this and many other cases is the initiation of a singular concept that marks the relevance of the object rather than the ”baptism” of the object (and of its concept) with a proper name.

According to this second meaning of uniqueness, I possess a singular concept of my dog Fido not only because I need to distinguish him from all the other similar dogs, but also because I want to remark the significance that this entity has to me. We refer to this property of singular concepts as *conceptual relevance*. Even though we noted that there are significant objects that we do not mark with proper names, there are many evidences that suggest that (at least for individuals of some kinds, like for example human beings) our distinguishing of entities according to significance is largely reflected linguistically in our practices of proper name-giving and name-use. We will discuss some of these evidences in section 3.4. Many relevant examples reported in section 3.4 have been suggested and extensively discussed by Jeshion in her paper about the significance of proper names [115] .

Another important distinction is between *singular concepts* and *object files*.

In the psychology of vision, several authors have hypothesized that the visual system uses temporary “object files” for tracking and identifying objects. For example, Kahneman and Treisman [120, 121] suggest that the main result of processing visually a particular scene is to construct a set of separate (visual) files, whose function is to store information about objects in the visible scene. An object file is responsible for the perceived continuity of the seen object. We will discuss in more details the characteristics of the object file system in section 3.1. Here, we simply contrast the notion of singular concept with that of object file to highlight the characteristics of the first kind of representation.

The main difference between the two kinds of representation can be understood quoting Kahneman et al. [121]. “We proposed an account of object perception as the process of setting up and utilizing temporary “episodic” representations of real world objects, which we call *object files*. Object files are separate from the representations stored in a long-term recognition network, which are used in identifying and classifying objects” (p.176).

An object file is therefore a *short-term representation* that allows the visual tracking of an individual in a perceptual field. It is a temporary representation which is addressed by its location at a particular time, not by any feature or identifying label and within which successive states of an object are linked and integrated, on the basis of spatio-temporal information. On the contrary, a singular concept is a *long-term representation* that allows long-term identification and entity tracking across lapses of attention, sleep, and other perceptual interruptions.

The notion of singular concept is more close to that of *mental file* proposed in theories of reference and singular thought [181, 7, 131, 116] in philosophy of mind and language. A mental file is a representation which allows the possessor of the file to store information about some thing, associated with some way of designating that thing. As a singular concept, a mental file is characterized by a particular relation that links it to some particular object in the world that the file is about. By virtue of this relation, the possessor of the file can think or speak directly of the object which is the referent of the file. Therefore, the notion of singular concept is used to explain the identification process, whereas that of mental file is used to explain mechanisms of direct reference in language and singular thought (i.e. thought about a particular individual). A mental file has a content which represents what properties the file’s possessor believes, intends or desires the referent of the file to have. In other words an agent’s mental files on objects capture that agent’s cognitive perspective on the world at that time.

## 2.2 Functional Dynamics of Singular Concepts

Adopting a feature-based approach, singular concepts can be represented as organized structures of semantic features which store our information about the individuals they are about. However, singular concepts are not mere long-term storages for entity-specific information but are indeed specialized structures for identifying and tracking unique individuals. Singular concepts are the core structures of singular cognition. We can characterize functionally singular concepts to serve two primary functions.

1. They constitute a cognitive system that allows an agent to identify and track unique entities across time and change. Adopting a term used by Jeshion in [116], we call this function *Identity Function*.
2. They serve as vehicles for bundling together an agent’s fund of information about a particular entity, providing an economical and efficient means of sorting, retrieving, and adding information on the particular individual. We refer to this function as *Storage Function*<sup>3</sup>.

We now describe in more details the functioning dynamics underlying these functions.

**1) Identity Function.**<sup>4</sup> When identifying an entity we access stored information (both perceptual and semantic) to decide which, if any, previously encountered entities corresponds to the entity presently encountered. Identification is a process across time. At the first encounter with an individual entity a singular representation, i.e. a singular concept in our framework, about that individual is initiated and different kinds of information are registered in it. For example, when I first meet a person I can register in my memorial representation structural aspects about her physical appearance (e.g. face, voice, body shape) and semantic information including biographical knowledge and her proper name. If I meet that person at a second time, visual, auditory or verbal inputs are processed, leading to the formation of a temporary description that is compared with all the representations stored in memory about known persons. If a match is found a singular concept is activated and the corresponding semantic information is available.

However, because individuals can change some of their properties while persisting as the same individuals, the cognitive system needs mechanisms to ensure

---

<sup>3</sup>Jeshion in [116] named this function “Bundling- function”

<sup>4</sup>Note that the idea that identity judgments should be understood in terms of their effects on the management of mental files (i.e. by initiating, updating, splitting and merging mental files) in the mind has been discussed by many philosophers of mind and language. See for example [131, 116, 181].



the identification of an individual as the same unique individual perceived or known at successive moments in time and across change.

We argue that these mechanisms are partly captured by functioning dynamics of singular concepts: initiation, updating, merging and separation.

- *Initiation*: a new singular concept is created when a new entity is assumed to come on the scene. For example, the first time that I know a new person, I store all the information about her in a new singular concept. Note that singular concepts, differently from object files, are concepts on individuals that we may or may not have directly perceived. We can have a singular concept for Napoleon or Pegasus even though we have never been in perceptual contact with these individuals.
- *Updating*: a singular concept is continually updated. The updating process ensures that the representation maintains its internal coherence and avoids the dissonance with the other beliefs of the subject. There are two aspects in the updating function. The first consists in adding new information to the representation. For instance, if a friend tells me something new about a known person, I simply update the corresponding singular concept adding the new information. The second aspect deals with revising information already stored about a known entity. If I come to know that a friend of mine moved to another city, I update the corresponding representation changing the specific information within the biographical knowledge about him.
- *Merging*: singular concepts are merged when the agent comes to identify two entities previously taken to be distinct. Imagine, for example, that you never met the sister of a friend. Nonetheless, you have a singular representation about her based on what you know from him. One day you met a girl at a party and a new singular concept is initiated about her in your memory. At a certain point during the conversation you understand that the person you have just met is in fact the sister of your friend. This means that now you have two singular representations about the same individual. In this case the two singular concepts need to be merged to create a unique representation which combines all the information stored in two original representations.
- *Splitting*: conceptual information is distributed in more than a single file when an individual previously thought to be one is thought to be more than one. For instance, going back to the previous example, if you think that your friend has only one sister, every time that he tells you about his sister you store the information into a unique mental representation.

But if you discover that in fact he has two sisters, you need to revise your memorial representations and eventually distribute different pieces of information into two different concepts.

The identification process implies that a unique referential link is maintained between an entity and its mental representation. It is important to note that there are two ways to activate *directly* a singular concept. The first is a *bottom-up way*. A perceptual stimulus activate several stages of processing that ultimately lead to activate the singular concept. The second is a *top-down* mechanism and concerns all the cases in which the individual cannot be perceived (or perceptual information is insufficient to go through the first way), but can be located or identified on the basis of information gathered by such sources as reasoning or communication. We argue that two different referential links are in play in these cases.

The first is mediated by structural information stored in the concept which is directly matched with the temporary structural representation of the perceptual stimulus. If the match is found the referential connection is established. Experimental evidence that we will describe in the course of this work shows that the bottom-up access to singular representations in memory is direct, that is, the initial point of contact between the perceptual stimulus of a unique distinguishable object and its memory representation is not mediated by high-level conceptual structures (i.e. general concepts). This means that having an individual representation of an object in memory (i.e. individual concept) shifts the entry point of recognition to the most subordinate level of the knowledge representation in memory, that is the unique level of identification.

The second is based on the mental counterpart of the main referential mechanism in language, i.e. the use of singular terms to refer to individual entities. Among the information contained in a mental file, there is typically that concerning the proper name of the individual the concept is about. This information has a different status within the singular concept compared to other information, as it is suggested by many studies which have shown that proper names are processed differently by the cognitive system than other kinds of information (see for example [217, 216, 98, 228]. We will discuss this issue in section 3.4). We assume that mental proper names serve as unique labels for singular concepts and are, typically, the prime means people use to create the referential link between an entity in the world and its singular representation in memory, whenever the singular concept can not be accessed directly via the perceptual way. Mental proper names are top-down modes of accessing the information stored in a singular concept.

Finally, there is an *indirect* way to activate a singular concept, i.e. by description. There are cases in which I have a singular representation of an individual, but I can not fix the reference of the representation by means of a unique label. I can have the singular concept for my sister's friend even though I don't know her name. As a consequence, the singular concept can not be labeled by means of a singular term. In this cases conceptual information can be used to fix the reference with a definite description (e.g. the sister of my friend Paolo) or with a set of attributes which can be used to single out the singular concept from others (e.g. the blond girl who lives in Trento and works at the post office). An interesting investigation about the different referential function of definite descriptions and proper names in initiating and merging singular representations in memory can be found in [3].

**2) Storage Function.** Insofar as singular concepts serve as vehicles for aggregating information that an agent has about a particular individual, they provide an economical and efficient means of sorting, retrieving, and adding information on a particular individual. This folder of information enables the agent to more easily access large units of information about particular objects and carry through inferences about such objects. We note that a proper storage function depends strictly by the same dynamics described above (i.e. initiation, update, merging and splitting) and therefore by a proper referential mechanism. Only if I identify the singular concept corresponding to the specific target entity which I'm processing, I can correctly manage the semantic information contained in it. In the present work our focus is mainly on the identity function of singular concept that is at the basis of a fundamental cognitive ability named *Singular Cognition*.

## 2.3 The problem of Singular Cognition

Singular cognition deals with two aspects of the functioning of singular concepts: identification and re-identification.

First, our conceptual system represents singular concepts and these concepts allow us to *identify* specific individuals as unique instances of these representations. Each singular concept corresponds to a specific individual and by means of a singular concept we can distinguish that specific individual from all others. If I am at the park with my dog Fido, my singular concept of Fido allows me to identify it as "my dog Fido" and distinguishes it from all other dogs. In this sense, singular concepts provide a means to guarantee the individuality or uniqueness of an entity.

Second, our conceptual system uses singular concepts to *re-identify* individuals over time and change. For example, if I identify that my dog Fido (the

unique referent of my singular concept about “Fido”) walked into a cabin, then a few minutes later, I again identify that Fido (as opposed to another dog) came out of the cabin. I ask myself: Was that my dog Fido going in and coming out, or was that two distinct dogs, one went in and the other came out? In this specific case, re-identification has consequences in whether our visual system interprets the event as having one or two specific dogs in the scene or whether I can use the same proper name, Fido, to call the dog which went in and came out of the cabin. However, in other cases re-identification may deal with long-lasting perceptual interruptions which may involve great changes in perceptual and non-perceptual facts about the entity which must be re-identified. Finally, there are cases in which re-identification may be performed exclusively on the basis of non-perceptual facts. In all these cases singular concepts provide a means to guarantee the identity of entities across time and change.

On the basis of this distinction, we define the problem of singular cognition as follows:

What are the conceptual capacities that are to be taken into consideration if one wants to explain how a cognitive agent can perform singular cognition, i.e. the identification and re-identification of an object as the same unique object at successive moments in time?

We believe that understanding the conceptual mechanisms underlying singular cognition is of fundamental importance to understand how people interact effectively with the entities relevant for their own existence. In particular we can summarize three main aspects of this interaction which are mediated by singular cognition. Singular cognition of an object  $o$  is necessary for:

1. for the acquisition, the rapid access and retrieval of specific knowledge bearing strictly on  $o$ ;
2. for maintaining a consistent representation of  $o$  across time and change;
3. for performing actions and having reactions that must be directed to that specific individual  $o$ . This includes, for example, an adequate use of an artifact, an escape reaction in response to stimuli indicative of danger or threat, an emotional reaction directed to a fiancé, a child or a friend.

Part I

Background and State of  
the Art



## Chapter 3

# Uniqueness in Cognitive Models

Many cognitive models assume the existence of mechanisms which ensure the identity tracking of unique individuals across time and situations.

There are at least two distinct representational systems underlying this fundamental aspect of human cognition. The first is perceptual and has been largely studied in the context of visual perception and infant cognition, exploring the principles by which the visual system segments the visual input in discrete objects and bind individual views of objects into dynamic representations which persist across time, motion, featural change, and interruptions.

The second system is conceptual and deals with higher level information that comes into play when an object is fully identified as an instance of a conceptual representation in memory. This system comes in play, for example, when an object is identified as a known individual both in presence and in absence of perceptual information.

In this chapter we review the cognitive models that have addressed the problem of object identity from these two different perspectives.

### 3.1 Singular Representations and Reference in Psychology of Vision and Attention

In the psychology of vision, several authors have hypothesized that the visual system uses direct mechanisms of individuation and reference which allow a cognitive agent to trace a perceptual stimulus in the visual field.

This is a fundamental process in visual cognition since our visual world is filled with objects that constantly change their position or appearance. The

shape, size, and position of an object on the retina change every time the object moves or we move our eyes. Yet despite these constant stimulus changes, objects in motion maintain continuity; likewise, objects are seen as continuous when viewed across saccades or during temporary occlusions, even though much of their appearance may change.

The visual system is therefore endowed with mechanisms which guarantee the maintenance of the perceived continuity of objects as they move, change, or momentarily disappear, ensuring the maintenance of what we call *perceptual identity*. It is worth to contrast *perceptual identity* with *conceptual identity*.

In *perceptual identity*, a stimulus retains its identity and continuity independently from the activation of its long-term representation in semantic memory. I can track an object in the visual field and perceive it as the same persisting object, even though I'm not able to identify it. Take the case of an observer who looks at an object moving slowly in the night sky. In the darkness of the night, the observer looking at the object may be not sure whether the object is an airplane or a falling star. Nonetheless, he is able to access the "perceptible sameness" of the object, without grasping the identity of the object. On the contrary, *conceptual identity* depends on a succession of states of activation of units in semantic memory which leads to recognize and fully identify an individual as the same individual that has been identified at another time and situation. If I meet an old friend after a long time, I may be able to identify him as the same individual I last saw ten years before, in spite of substantial changes in perceptual appearance. I track my friend on the basis of non-perceptual facts because I'm able to access the "conceptual sameness" of that individual.

We argue that singular concepts are the critical representations to ensure conceptual identity.

In this section, we review the mechanisms which have been proposed in psychology of vision to support perceptual identity and we contrast these mechanisms with those involved in conceptual identity through singular concepts.

### 3.1.1 Object Files

At present, it is unclear how the visual system preserves object continuity despite stimulus changes. One possible explanation, known as *object file theory*, has been proposed by Treisman and her colleagues [120, 121, 248].

According to this theory, when attention is directed to an object in the visual field, a temporary representation of that object, i.e. an *object file*, is created. This file is an episodic, visual representation which store and update information about the object it represents and it is kept open so long as its object is in view and may be discarded shortly thereafter.



Object files are defined as perceptual units into which a scene is parsed becoming the potential objects of attention. Apart from being vehicles to bind features on which the attention can be allocated, object files are thought to bind successive states of an object over time, updating their representations as the objects move and change. It is argued that this mechanism is at the basis of the capability of the visual system to restore continuity that has been briefly broken in the stream of sensory inputs (e.g. during saccadic eye movements or temporary occlusion).

Object continuity is maintained through a process that consists of three operations: *correspondence*, *reviewing*, and *impletion*. 1) A correspondence operation determines, for each object in the visual scene whether it is “new” or whether it is an object recently perceived, now at a different location. This determination is based on low-frequency spatiotemporal information; features such as shape, color, or identity are irrelevant to the correspondence problem; 2) A reviewing process retrieves the content of the initial object and compares it with the characteristics of the object in the current scene. If there is a match, object continuity holds. If the appearance of the object in the current scene is inconsistent with the previous object file contents, however, the object file must either be modified, or discarded and replaced with a new object file. 3) Finally, impletion operations use current and reviewed information to establish a link between previous and current object files by creating the appearance of change or motion in the scene.

An interesting question about object files is what kind of information is included in an object file.

To answer this question Kahneman et al. [120] introduced an experimental paradigm known as *object-reviewing paradigm*.

In the initial object-reviewing experiments [121], observers viewed a “pre-view” display that contained two or more objects with a different letter placed in each. The letters then disappeared and the objects moved to new locations. Once the objects stopped, a single “probe” letter then appeared in one of the objects, and the observers simply named it aloud. The probe could be one of the initial preview letters (on “match” trials) or it could be novel to the trial (on “no-match” trials). Further, on match trials, the probe letter could reappear on the same object in which it had been previewed (on “congruent” trials) or on a different object (on “incongruent” trials).

Using this paradigm, Kahneman et al. [121] reported that naming latencies were longer when the target letter was a repetition of the preview letter from the opposite object (i.e. incongruent trials) than when it was a repetition of the preview letter in the same object (i.e. congruent trials) - an effect termed *object-specific preview benefit* (OSPB). This result suggested that object identity may

be included in an object file. In particular, Treisman and her colleagues have suggested that all information that defines an object is included in its object file, including identity and meaning. In particular to mediate recognition, the sensory description in the object file is compared to stored representations of known objects. If a match is found the identification of the object is entered in the file, together with information predicting other characteristics, its likely behavior and the responses it should appropriately evoke, both affective and cognitive.

Other researchers have used this basic paradigm to argue for the exclusion of certain object characteristics from object files. For example, Henderson [4] changed the type font of a single letter between successive displays and found that the change did not eliminate object-specific effects. This suggests that information about exact physical form may not be included in an object file.

Again, the results reported by Gordon et al. [89] suggested that information about the identity of objects is stored in object files, but at least three types of semantic information (related concepts, semantic features, and category membership) are not. The same authors found in another study that a concept can be represented regardless of its medium (e.g. the abstract identity “fish” persists despite being previewed as a word and probed as a line drawing [88]).

These results seem indicate that object files include object identity and abstract information.

However, in a recent study Mitroff [163] performed a object-reviewing experiment with novel face images as stimuli and found that object files can store not only abstracted information about object types, but also specific visual features of individual object tokens.

From these premises we note some important differences between the object file system and that of singular concepts. First of all, object files are separate from the representations stored in long term memory which are used to classifying and identifying objects, i.e. general and singular concepts respectively. In particular, in contrast to singular concepts which serve as storage mechanism of long-term identification networks, object files are temporary representations which are addressed by their location at a particular time, not by any feature or identifying label and within which successive states of an object are linked and integrated.

Unlike long-term object identification, where surface features may be used for bottom-up identification (e.g. recognizing a friend across a crowded auditorium), object files are mid-level visual representations (mid-level because they fall between low level sensory processing and high level placement into conceptual representations) which operate in online visual processing tracking objects on the basis of spatio-temporal information, i.e. how and where objects move

rather than how they look like.

The identity of a changing object is carried by the assignment of information about its successive states to the same file, rather than by its name or by properties. This assignment leads to update and review object file content, but the singular content is not used for object tracking.

Therefore, object files can be considered the placeholders for that of updating system (or updating state with singular content).

Differently from singular concepts, object files can not be accessed by means of top down processes of identification and they are responsible for the perceived continuity of the seen object without need to access to semantic information about it.

In conclusion, object files are cognitive representations which have been proposed as core mechanism of singular perception - i.e. the incremental perception and tracking of an object as the same unique, or distinct and numerically identical object - whereas singular concepts are here suggested to be the fundamental mechanism of singular cognition - i.e., the identification of an object as the same unique object identified at successive moments in time, even in absence of perceptual information.

### 3.1.2 Visual Indexes

Attempting to answer the question of how the world is connected to our visual representations, Pylyshyn has proposed a theory of vision that assumes the existence of a special kind of direct connection between elements of a visual representation and certain token elements in the visual field [187].

This connection is unmediated by an encoding of properties of the individual tokens involved and implies a sort of direct link between a perceptual system and an object in a scene. Like natural language demonstrative (i.e. deictic words like *this* or *that*), this direct connection allows entities to be referred to without represent them “under a description” (i.e., without representing them as members of some categories).

Such a preconceptual connection is ensured by a mechanism of visual indexes or visual demonstratives (or “FINSTs”, from “FINgers of INSTantiation”) that picks out individual objects from the rest of the visual field and allows to maintain and track the identity of these objects qua individuals despite changes in the individual’s properties. As we noted above, FINSTs are more similar to demonstrative than to proper names in natural language. This is because they ensure a direct reference to a particular individual but this reference relation ceases to exist when the referent is no longer in view. This way of reference is also “preconceptual” since it allows to refer to things in visual scenes regardless

to their category membership and pick out them directly by a mechanism that works like a demonstrative. To put the point in other words, if visual indexes provide a mechanism for reference to distal items, this is not a kind of reference by description, but reference constituted by some sort of causal connection between the object in the world and its visual representation in the cognitive system.

This mechanism of reference provides a means by which the cognitive system can pick out a small number of individuals in a visual scene, keep track of them and further examine them in order to encode their properties, to move the focal attention to them or to carry out a motor command towards them.

Pylyshyn [187] suggested that conceptual or descriptive representations are insufficient as the sole form of visual representation. According to this view, If we could refer to the elements of a visual scene only in terms of their category membership, our concepts would always be related to other concepts and would never be grounded in experience. Moreover there are two general problems raised by the description view of visual representations. The first is that there are an unlimited number of entities in the world that can belong to any particular category or satisfy a particular description. As a consequence, reference by category or description may be inadequate to refer to a unique individual among many similar ones in the visual field. Secondly, the visual system needs to be able to individuate a particular object in the visual scene and track it as a particular enduring individual in spite of its property changes. A visual tracking by description would be extremely expensive because the description would have to be continuously updated with the changes of the object (such as changes in visible surfaces or spatio-temporal location).

Some empirical reasons have been proposed to motivate the existence of primitive indexing mechanisms as a possible solution to the previously mentioned problems.

First of all the mapping from the world to our visual representation is not built up in one step but incrementally (for example scanning attention and/or one's gaze). This implies that all the information about a particular token acquired across different periods of time should be associated to the same individual object. A descriptive approach to this problem would need a description that is unique to the individual in question, say "the object  $x$  that has the property  $P$ ", where  $P$  uniquely picks out that particular object. In order to add new information about the object, you need to add a new predicate  $Q$  (say "the object  $x$  has the property  $Q$ "). This way of adding information also requires an identity assertion that specifies that the two properties refer to the same object ( $P(x) \wedge Q(y)$  and  $x \equiv y$ ). This way of representing and updating the information about an individual object presents a main problem. In order to

add information about an object  $x$  in the visual field, first you need to recall the description under which  $x$  was last encoded and then add a new predicate to the original description. In other words, every time you gain new information about a certain object  $x$  (say a property  $Q$ ), you need to go back to the previous representation of it and detect that the object that you noted as having a new property  $Q$  also had a previous property  $P$  which uniquely identifies it. After having established the identity, you need to update the representation and describe the object as the referent of a new description that uses the conjunction of the two property ( $P$  and  $Q$ ). This backward process from an individual to its previous representation seems implausible and antieconomical.

Another solution to the problem of updating a representation upon noticing a new property  $Q$ , invokes a direct mechanism of reference. This mechanism does not need to locate a representation of an individual with certain properties, but rather needs the direct link to the very individual on which the new property  $Q$  has been detected, regardless of the properties you have already encoded at that point of time. The author suggests that this mechanism is mediated by some functional equivalent of demonstrative reference. Adding a new property  $Q$  to a representation of an individual object  $x$  requires adding a new predicate  $Q(x)$  where  $x$  is the object directly picked out by the demonstrative indexing mechanism.

A second reason is that there are many properties that are extracted and encoded like relational properties (e.g.,  $Inside(X, C)$ ) which apply over a number of particular individuals. In order to apply these properties we need to specify which objects are involved in the relation independently of what properties these individuals have. The visual system must adopt a mechanism that uses something different from descriptive information in order to track individual objects and their relations in the visual field. Like proper names or demonstratives in natural language, this mechanism uses visual indexes that uniquely pick out particular individuals. These indexes may be used as labels or names that refer directly and a-conceptually to individuals. This means that we have a way to individuate and track individual objects in a scene even when they change their properties or location.

Finally many evidences (see [212] for a review) support the assumption that a property is detected and encoded by the visual system, not just as a property existing in the visual field, but as the property of a specific perceived object. This object-based encoding must be guaranteed by a direct mechanism of reference which allows that properties are always detected as belonging to an object.

We briefly summarize the main characteristics of the FINSTs system. The FINST indexing mechanism is the way of reference that the early visual system uses to pick out and track individuals in a scene without recurring to

top-down conceptual descriptions. This mechanism is *preattentive*, *preconceptual* and *bottom-up*. The individuals picked out by this mechanism are named *primitive visual objects*. The basic idea of the FINST indexing mechanism is to provide a mechanism to link these primitive visual objects (by means of FINST indexes) to certain conceptual structures (which we may think of, in our framework, as singular concepts in Long Term Memory). This connection is purely causal and stimulus-driven without cognitive or conceptual intervention. An individual object “grabs” the indexing early vision mechanism and thus initiates a FINST. The number of FINSTs that can be activated at a single time is limited (by means of this mechanism we can track only four or five individual objects.) By virtue of this causal connection, the conceptual system can refer to any of a small number of primitive visible object. It can, for example, move focal attention to them, evaluate visual predicates and finally predicate something about them. It is interesting to note that claiming that the indexing process is preconceptual is not to claim that the assignment and maintenance of indexes does not involve the properties of objects. Clearly indexes get assigned because objects in question possess certain properties rather than other properties. What is claimed is that the encoding of these properties is not necessary to the cognitive system to assign and track an index. Without preconceptual reference we would not be able to decide that a particular description  $D$  was satisfied by a particular object  $D$  and consequently we could not make judgments about nor to decide to act upon a particular individual.

It is interesting to contrast visual index theory with the object file framework and with our notion of singular concept. The FINSTs theory is very close to the object file theory of Kahneman et al. [121] described in 3.1.1, even though the latter was more focused on the memory content of the information associated with the object in memory. This focus is also confirmed by a lot of research on what object-related information is encoded in an object file [262, 89, 163].

Kahneman et al. suggested the relation between object files and visual indexes, when they write “We might think of [a visual index] as the initial spatio-temporal label that is entered in the object file and that is used to address it . . . [A] FINST might be the initial phase of a simple object file before any features have been attached to it” (p. 216).

Because of this difference in focus, research on visual indexes has more concentrated with the nature of the reference mechanism that allows cognition to refer directly and track objects, whereas object file theory has more focused on the question of which features of the object are encoded in memory.

Both systems concern temporary representation of objects and have been proposed to address the problem of perceptual (i.e. visual) identity. While visual index theory emphasizes the mechanism that connects representations with the

objects they are about, object file theory is more focused on the question of which features of the object are encoded in memory. Object files store temporary “episodic” representations of objects in a recent visual field that is updated through alterations in the perceptual situation. FINSTs are the vehicles by which object files represent the objects that they store information about. What this means is that, while object files collect information about objects, this information is not used to determine which individual it is associated with. It is the FINSTs system that creates the bridge between the object file and the individual that the file is about.

Differently from object files and singular concepts, visual indexes are deictic non conceptual mechanisms of direct reference in vision. While they provide the connection between a concept and an object in the world, the conceptual representation of the object is stored temporarily in object files or permanently in singular concepts. This means that FINSTs and object files are mechanisms that allow to keep the visual identity of objects, whereas singular concepts are conceptual mechanisms of reference used to track the conceptual identity of objects across lapses of attentions, sleep and different kinds of perceptual interruption which can not be dealt by the cognitive system with temporary mechanisms of reference and representation.

Prima facie, there is a striking similarity between the psychological theories about visual indexes and object files on one hand and our notion of singular concept on the other hand. Both accounts 1) are object-centered approaches of mental reference and 2) share some primitive functioning notions about the organization of the representation of individuals (e.g. the need of a direct mechanism of reference or updating functions). Nonetheless, the roots of the tradition in psychology of vision differs from our approach in at least one important respect: in psychology of vision, the theoretical constructs of visual indexes and object files appear as a rather non-conceptualist solution to the problem of tracing the perceptual identity of objects across time and change because it refers to a temporary visual representation which can track a persevering object in the visual field without the use of sophisticated conceptual or descriptive contents. On the contrary, singular concepts provide a conceptual solution to the same problem and are at the core of a storage mechanism of long-term identification networks which involve more complex conceptual representations of individual entities.

### **3.1.3 Object Indexes and Object Concepts**

From a different perspective, developmental studies have faced the problem of how infants establish representations of individuated objects and track them

through time, space and occlusion.

Recently, many have suggested that the studies of object representations in infancy involve similar problems (and plausibly the same psychological mechanisms) to those reported in the mid-level object attention studies in adults (see for example, [135, 41]). In a paper published in 1998, Leslie [135] reports a series of consideration in favor of this hypothesis and presents a model of object representation that underlines the main similarities within both the literatures. The key notion of this model is that of “sticky” object index, a mechanism of selective attention that points directly at a physical object in a location. Just like a FINST in the Pylyshyn’s model, an object index does not represent any of the properties of the object which it points to. The indexing mechanism forms the basis for the infant’s object concept because it is involved in object individuation, identification and enumeration of physical objects. An object index has a certain number of properties that closely recall some characteristics of both of models described above (that of FINSTs and that of object files). First of all, an object index is a mental token that functions like an abstract pointer to an object in the world.

Second, an index does not inherently represent any of the properties of the object indexed. However, this information can be bound with the index and can be used in the identification process.

Third, object indexing is a mechanism of selective attention and presents resource limits. This means that only a limited number of indexes can be associated to specific objects in a scene (not more than four).

Forth, indexes are assigned to objects primarily by location but they are not linked to the location itself but to the object in that specific location. Moreover, property information eventually bounded with the index can be used for the index assignment when location information is unavailable or ambiguous.

Finally, there are some basic principles that control the allocation mechanism.

A distinct object can be assigned only with a single index and, when assigned, the index sticks to its target through space. This mechanism provide immediate access to the object’s location even though the object is in motion or it moves behind an occluder. In the last case the index points to an approximate location behind the occluder.

Distinct indexes are assigned to objects that occupy different locations in space at the same time.

Finally, indexes can be reused and reassigned to different objects when they are disconnected from their previous targets. The index reassignment is necessary because only a small number of objects can be indexed simultaneously.

In order to understand the functioning of the index system, it’s important



to note the difference between two distinct processes, object individuation and object identification. *Individuation* establishes the notion of “single object” and “more than a single object”, whereas *identification* established the notion of “self-same one”.

The theory assumes that object individuation is primarily determined on the basis of the locations objects occupy and not on the basis of the features they have. However, featural information is assumed to be integrated at a later stage to the early object representation. Featural information is also necessary in those situations in which spatiotemporal information is absent or ambiguous. For example, if a cup and a ball take turns appearing from an occluder and disappearing behind it, we judge that there are two objects in the scene rather than one whose features change. This judgment is based on featural differences because the two objects are never seen simultaneously occupy different locations. Featural information influences the indexing process in two distinct ways: *individuation-by-features* and *identification-by-features*. In the first case, the system simply registers whether or not salient new features have appeared (feature detection). The second type of processing encodes specific featural information that is bound to the early representation after an index is assigned (feature identification).

The first type of output suffices to count how many objects are present in a scene, but the latter information is required in order to identify what objects are in play.

The results of many studies using the violation of expectancy looking time paradigm have been interpreted in the light of indexing model, showing that the mature indexing system can assign indexes either by location or by features. However, a body of findings provide evidence that the object individuation process undergoes many changes and that a complete individuation-by-features is not available by the age of 12 months [269].

## 3.2 Singular Representations in Models of Face Recognition and Naming

In section 3.1 we have discussed models of vision and attention which propose direct mechanisms of individuation and reference to explain how a perceiver can perform the perceptual individuation or identification of an object as the same unique object perceived at successive moments in time (i.e. singular perception). All the models that we have reviewed above share a non-conceptual approach to the problem of singular perception since they are based on the idea that sensory-motor capacities or perceptual contents, make it possible for a perceiver to latch

on to, or to track a target  $x$  as being the same target without the help of complex conceptual or descriptive capacities.

Such perceptual capacities must be able to perform anchoring of the perceiver onto the object  $x$  and provide perceptual reference to  $x$ , regardless of the fact that the object is fully identified as a unique individual (e.g. the object  $x$  is my dog Fido) or as a member of a category (e.g. the object  $x$  is a dog). In other words, instead of tracking  $x$  over time and space on the basis of the understanding that  $x$  is a known individual which has a unique representation in memory or that  $x$  is an exemplar of a learned category, non-conceptual approaches anchor the perceiver on to  $x$  without the mediation of an elaborate understanding of the “conceptual identity” of  $x$ .

Moreover, these models deal with the problem to explain how a perceiver trace the identity of an object when the perceiver is in “perceptual acquaintance” with the object or the perceptual acquaintance is only temporary interrupted.

However, when we have to recognize and track individual objects over long-term interruptions or changes in perceptual properties or even in absence of perceptual inputs, high level information about identity (i.e. conceptual representations) must come into play.

In particular, unique high level representations about objects are involved every time a unique individual is recognized as a known individual and individual-specific information is retrieved about it.

In the literature on object recognition, almost exclusive attention has been given to a special kind of unique individual entities, i.e. person, and very few studies have investigated the recognition of other kinds of unique entities<sup>1</sup>.

Many models of face recognition and naming assume the existence of unique representations of individuals in memory and the way in which conceptual and name codes of familiar faces are accessed from perceptual input is a matter of considerable current debate in cognitive research. In this section we review the major cognitive models that have addressed this issue and have inspired research questions and experimental paradigms useful for the present research.

A first comprehensive model of face recognition was proposed by Bruce and Young [35], which assumes that access to face names is the last step in a serially arranged sequence of processing stages. The model proposes three main representational stages: 1) a *recognition stage*, which involves a set of structural and view-independent long-term representations (face recognition units or FRUs); 2) a *semantic stage*, which permits the activation of permanent, person-specific knowledge about the recognized person; and 3) a *name retrieval stage* which allows the retrieval of the proper name of the person.

According to this model, perception of a familiar face activates structural

---

<sup>1</sup>Some exceptions are described in Chapter 4.

and view-independent long-term representations (FRUs). Each known face was assumed to be represented by a FRU, the activation of which permits a familiarity decision (i.e. the decision that a face is “known” ). These FRUs are linked to amodal person identity nodes (PINs) which contain semantic-biographical information concerning known persons, such as occupation, hobbies, date of birth, etc. In a final step, name nodes are accessed from their corresponding PINs. Bruce and Young claimed that processing occurs in the fixed (and immutable) order from FRUs to PINs and then from PINs to retrieval of name codes, and it was further assumed that processing must be completed at one stage before it starts at the next. As a consequence, the model assumes that naming is necessarily semantically mediated and names are harder to retrieve than other person-specific information because names are stored in a separate and final component, which may be accessed from faces only via semantic information.

However, the Bruce and Young’s model does not have a route for the production as opposed to the retrieval of names or other personal information about people. There are many possible output systems which could be recruited following the initial retrieval of person-specific information or the corresponding proper name. For instance, one might be shown a picture of a face and be required to pronounce the person’s name, to write it or to press a button in a laboratory experiment. Moreover the model is also incomplete in the sense that it shows no route by which a input names can access to personal information.

In order to account for these processes, Valentine et al. [250, 251] proposed a functional model of face, name and word recognition which is an extension of the original model by Bruce and Young. The author proposed a further stage of word recognition units (WRUs). There is a WRU for each known word and this unit becomes active as a result of input from any recognizable instantiation of the word. Those words that are names activate a new set of units named name recognition units (NRUs). These units are thought to be analogous to FRUs, i.e. there is a NRU for each familiar person. The activation flows from NRUs directly to PINs. However, unlike FRUs, these units have direct access to the lexical output codes. The connection between NRUs and PINs serves to link the conceptual system with lexical representations. Access from a face representation to a person’s name can only be achieved by this single link from the PIN to the NRU that represents the name. Like in the Bruce and Young model, in the model of Valentine et al. PINs are units which store semantic information about people and separate semantic stores are assumed for information about people and words.

An alternative architecture can be found in the interactive activation and competition (IAC) model developed by Bruce and Burton [38, 37] which is based on the architecture described by McClelland and Rumelhart [152].

This model comprises three sets of units of processing: Face Recognition Units (FRUs), Person Identity Nodes (PINs) and Semantic Information Units (SIUs). The units are organized into pools such that the units within a pool are connected to each other with inhibitory links. The links between units belonging to different pools are excitatory. All the links are bidirectional and have equal strength in each directions. Furthermore, all the excitatory links have equal strength and the same is for the inhibitory links.

For each face there is a single face unit which becomes active. FRUs are connected to corresponding PINs, that are multimodal units receiving inputs also from other systems (e.g., a PIN is activated by read names, voice and so on). PINs are cross domain gateways to semantic information stored in SIUs and signal familiarity. When a certain threshold is crossed, the face is recognized as familiar and the PIN leads the activation to the corresponding SIUs. In turn, the activation of a SIU above its threshold corresponds to the retrieval of the corresponding personal information encoded into it.

We note that the Burton and Bruce’s model differs from the Bruce and Young’s model in some important respects.

First of all, as in the Bruce and Young’s model, PINs are activated from their corresponding FRUs, but differently from the Bruce and Young’s model, these PINs do not contain identity-specific semantic information but they permit access to it. PINs merely serve as modality-free interfaces between FRUs on the one hand, and both semantic-biographical information and names on the other hand.

Secondly, units in the SIUs pool are supposed to specifically code person-specific knowledge about people, as well as the names of these persons. Therefore, in the Bruce and Young’s model, name retrieval takes place in a separate processing stage that follows, and is contingent upon, the retrieval of semantic information about the person. In contrast, Burton and Bruce proposed that names and semantic information can be accessed in parallel. Hence, the assumption of conceptual mediation prevalent in the serial model of face naming proposed by Bruce and Young is abandoned.

Therefore, the two models make different predictions about whether face name retrieval is subject to semantic context effects. The serial account assumes that face naming mandatorily proceeds from face recognition to name retrieval via semantic representations and therefore semantically related primes should induce priming effects. The parallel account, on the other hand, does not necessarily predict semantic effects in face naming, as names can be accessed independently from semantic codes.

However, the activation mechanism at the core of the model proposed by Burton and Bruce predicts more complex interactions between semantic and

identity units due to the back propagation of activation in the network.

The model assumes that when a particular PIN is activated, activation from the PIN flows to the SIUs that are connected to it. Some activation flows back from these SIUs to PINs that share semantic features with the original person, taking activation in any such PIN above its resting level. For example, suppose input is given to the FRU of Barack Obama. Activation flows to Barack Obama's PIN, which in turn activates the SIUs with which he is associated (e.g., President, Michelle Obama, USA etc.). As George Bush's PIN is also connected to many of the same SIUs, activation spreads back to George Bush's PIN taking it above its resting level. The level of this "above resting activation" depends on how many semantic features are shared. If at this point input is given to George Bush's FRU, activation will flow to his PIN, which will reach threshold faster than had it started at resting level, and this is the basis of the facilitatory effect.

Note that this architecture implicitly assumes a form of categorical organization between the identity nodes of the network. Since each SIU is connected to the PINs of persons who share the same attribute (e.g. the SIUs representing occupation information are connected to all persons with the same occupation), when a familiar face is presented, activation can spread back to the representations (PINs) of other persons also linked to the same SIUs, e.g., persons with the same occupation. In this way the shared category functions as an organizing category of person-specific nodes, and the IAC model predicts that categorical priming should be observed between two persons sharing a common category (e.g. occupation). This assumption of the model has been largely investigated in priming experiments and has motivated many researches which explored how semantic knowledge for people is stored in long-term memory (see for example [9, 9, 236]). We will discuss in more details these studies in chapter 7, because they have been the starting point for methods and research questions which have inspired the part of our work concerning how singular concepts are organized and accessed and how these representations are inter-linked with those of other individual things.

The Burton and Bruce's model provides also a different explanation, compared to the Bruce and Young's model, about why names are difficult to retrieve compared to other semantic information. Quoting the authors, "while we know many teachers, many Americans and many politicians, we typically know only one Margaret Thatcher". Therefore, units in the pool of SIUs that represent a person's full name are connected to only one PIN, but units representing semantic information are linked to many PINs. If a unit of FRUs pool crosses the threshold, the activation passes to the corresponding PIN and in turn to the SIUs connected to the specif person. Because of the nature of connections (that

are bidirectional), SIUs that are unique in the network are activated slowly than SIUs that are connected to many PINs.

For example, if a picture of Barack Obama is seen, the SIU that codes that he is a politician is activated and the activation is back-propagated to all the PINs associated with other politicians. This activation passes again from these PINs to connected SIUs. In this way SIUs that are shared by many PINs receive more activation than unique SIUs. As a consequence, any “unique” semantic information that is known about a person (like for example the proper name) should be more difficult to retrieve.

Brédart et al. [36] have proposed an alternative architecture for the interactive activation model in which descriptive properties are represented in separate pools of units for each semantic domain of information and in which names are represented by a separate pool of lexical output units. In particular there are two main problems with the original model that the new version tries to resolve.

First of all, the Burton and Bruce model predicts that the more facts you know about a familiar person, the slower you should be to retrieve any of those facts including the name (fan effect). This is because all identity-specific semantic information is represented by SIUs within the same pool of units and inhibitory links are assumed to connect these units. If many SIUs are activated by the same PIN, the amount of inhibition between the units within the same pool increases. The empirical evidence used by Burton and Bruce to give support to the validity of their model is controversial. In particular, the authors argue that their prediction is supported by the reverse frequency effect in retrieval failures for names - that is the fact that retrieval failures are reported much more often for names that are rated as familiar than for names that are rated as not very familiar. However the reverse frequency effect is not always reported. Brédart et al. [36] have conducted a study to evaluate directly the relationship between the number of properties known about people and the retrieval of those people names. The results of this study show that naming the face of a person about whom we know many pieces of information is faster than naming a person about whom we know few pieces of information, although the two sets of items were equated for face familiarity. The results are in the opposite direction of what predicted by Burton et al.’s model.

The second problem with the Burton and Bruce’s model, is that the storage of names and semantic information within the same pool of units is inconsistent with models of speech production. In particular it’s not clear the status of the SIUs representing names (prelinguistic units or lexical?) and why this kind of information should be stored alongside semantic information.

On the basis of these considerations, the authors proposed two main modifications to the original model.

First, storage of names and descriptive properties are separated into different pools of units. Names are represented by lexical units in the output lexicon.

Second, personal information is clustered into semantic sub-domains.

Like in the original model, units within a pool are connected each other with inhibitory links and units may be connected across pools by excitatory links. All connections are bidirectional.

A first pool of units (PINs) contains token markers, one for each familiar individual. These units are connected to semantic information units (SIUs) representing personal information about known persons. Each PIN is also connected to one Lexical Output Unit (LOU) representing lexical access in production of their names.

With a series of simulations, the authors provide an interesting set of evidences. First of all, the network is able to exhibit properties consistent with mental chronometry as effectively as Burton and Bruce's model. In particular, the fact that semantic information is accessed more rapidly than lexical information is confirmed by the pattern of mean activations of lexical and semantic units. It's never the case that a name unit reached the threshold activation before any of the SIUs associated with the same PIN. Moreover names rose slowly the maximum level of activation that SIUs except for SIUs representing unique properties.

A second simulation compared the effects of an impairment of lexical access (attenuating the PIN-LOU connection) on the retrieval of semantic information to the effects of the impairment of semantic access on the name retrieval (attenuating the PIN-SIU links). The results show that the first kind of manipulation do not prevent SIUs from reaching the threshold of activation, while the second alteration does prevent LOU from reaching the threshold.

Finally, the model confirms the prediction that the more properties are known about an individual the easier his or her name is retrieved.

The comparison of different models of person recognition and naming opens interesting questions about the structure of the semantic memory, the access and the retrieval of personal information of individuals. All the models agree that several stages are involved in the process of recognizing and accessing information about people. First of all, a visual, auditory or verbal input is processed, leading to the formation of a structural description that is compared with all the structural representations contained in modality specific units (respectively Face Recognition Units, Voice Recognition Units or Name Recognition Units). Secondly, modality specific units converge into Person Identity Units (PINs) allowing recognition of a particular person and activation of the corresponding semantic information. The third stage is the retrieval of biographical information associated to the specific individual and finally the process allows the

production of the person’s name. These models differ, however, in two important aspects concerning the locus in which familiarity feelings are generated and in which person-specific information is stored. Furthermore, controversies exist about the format in which biographical knowledge is represented.

Both in the Burton et al.’s and in the Brédart et al.’s models, a supra-modal level of PINs is responsible of the generation of familiarity feelings. At this level information from different modalities is combined in personal identity nodes that do not store semantic information, but provide a modality-free gateway to a single semantic system, where information about people is stored in an amodal format. From this respect, these models differ from the face identification model originally proposed by Bruce and Young which locates the locus of familiarity feelings at the level of recognition units where the structural description of a seen face is compared to the familiar faces stored in the FRUs. Moreover, the Bruce and Young’s model assumes that PINs store semantic information.

Apart from the differences between the face recognition models, what is relevant for the purposes of the present research is that all the models assume the existence of identity nodes in semantic memory which provide a mechanism of unique reference establishing a relation one-to-one between an individual in the world and its memorial representation in the cognitive system. Even though some models [35, 250] assume that identity nodes directly store person-specific information, while other models [38, 37, 36] represent this information in separate semantic nodes, the common idea is that the cognitive system use different structures to store general knowledge from those used to store individual-specific knowledge. This view is in line with our notion of singular concept, i.e. a cognitive representation of a unique individual which promotes recognition and identity judgments.

Another common aspect of these models is the special status of proper names among other person-specific information. This is in part due to the fact that these models have been influenced by a considerable body of evidence which support the view that the retrieval of proper names is in some way different from the retrieval of other personal information.

First of all, experiments have shown that people are slower to name familiar faces than they are to categorize the same faces by occupation. For example, when subjects are shown a face, person’s name is retrieved more slowly than other personal information such as occupation or nationality [273, 46].

In tasks that require subjects to learn face-occupation-name association they are generally showed more difficulties in learning people’s names than in learning semantic information about them [45], even when the words to learn are the same like in name-occupation homophones [158], (i.e, learning that somebody’s name is “Backer” is more difficult than learning that somebody is “a backer”).



This means that the name is more difficult to recall and this effect can not be attributed to differences in the phonological form or frequency to occurrence.

Retrieval of some personal information is possible without retrieval of a name but the converse has never been demonstrated. Diarists involved in studies of everyday difficulties in person recognition commonly reported incidents during which they are unable to retrieve a person's name while being able to remember a lot of personal facts about him [213]. They don't, however, remember a person's name without remembering his or her occupation [272].

The vulnerability of the retrieval of proper names is also seen in the tip-of-tongue phenomenon (TOT) in which a person is unable to produce a word although he is certain that the word is known. Many studies have shown that this phenomenon for proper names can be induced in the laboratory, showing pictures or presenting verbal descriptions of famous people [102]. The TOT state represents an impairment of phonological information in a name, not semantic information associated with it. This is also confirmed by the fact that cueing with the name's initials aids the resolution, while presenting other cues like pictures did not.

Further evidence for a distinction between retrieval of names and retrieval of biographical information can be found in neuropsychological literature. Several studies of anomic patients [69, 143] describe people who are unable to name familiar people while being able to access to relevant semantic information about them. No cases of patients showing the converse pattern have been reported in literature.

Many models of face recognition and naming explain these evidences assuming that names are stored separately from other kinds of personal information [35, 250, 36]. For instance hierarchical models such as the Bruce and Young model [35] posit a store for names that is functionally separate from the store for other personal information and which can only be accessed after that all personal information is retrieved. Although this suggestion is consistent with empirical evidences, it seems to be problematic. It is possible recognize Tom Cruise and recall his name, but be temporarily unable to remember that he was married with Nicole Kidman. IAC models like that of Burton et al. [37] do not separate the representation of proper names from that of other personal information, being stored in the same semantic units (SIUs), but they assume that proper names are hard to retrieve because they are unique and therefore they can not be pre-activated by the activation of other PINs. However, more recent models, like that of Bredart et al. [36] propose that storage of names and descriptive properties are separated into different pools of units but they are accessed in parallel instead of serially.

Our notion of singular concept is more similar to that of PIN in the model

of Bruce and Young [35] in that it assumes that a singular concept is an identity node which store individual-specific knowledge about an entity. Within a singular concept, as within a PIN there is all the semantic information we know about the represented entity. Note that in the IAC models the complete representation of an individual is given by the conjunction of relevant SIUs connected with the unique PIN which corresponds to the entity. This means that the knowledge is distributed in a network of nodes which are connected to many PINs.

However, our notion of singular concept shares some characteristics with the notion of semantic node (SIU) of the IAC model of [37] in that it does not separate the representation of the proper name from that of other personal information. This means that the mental representation of the proper name is part of the singular concept as it is all the information we know about the entity the concept is about. However, the mental proper name has a special status among other information in the singular concept, because it functions as a mental label that can be used as the mental counterpart of a singular term in language to uniquely refer to a mental representation in memory. The mental name is a sort of unique identifier for a mental representation as the proper name is a rigid designator for reference in language.

There is a third aspect about the organization of the semantic information of unique individuals in face recognition models which opens an interesting research question for the present study. These models posit that all the semantic information about a person is equally important in the semantic representation of that individual. In the hierarchical models this is confirmed by the fact that all the information stored in the PIN must be accessed before passing to the next stage of the recognition process and no claim is made about which piece of information is activated before another within the PIN. In the IAC models this assumption is manifested by the general architecture of the model. All the excitatory links as well as all the inhibitory links, have the same weight. This means that when a PIN is activated, the activation pass to all the SIUs connected to it with the same activation power. As a consequence, all the information about an entity is equally available for retrieval, unless some SIUs are pre-activated from other PINs.

We believe that this assumption of equality of importance between the semantic attributes of an individual is implausible since it is evident that some attributes about an entity can be accessed more rapidly than others even though they are part of the singular representation of that entity. Therefore we argue that a model of singular concept should account the differences between attributes in terms of relevance for the representation and its access. To this purpose, one of the goals of this work is to suggest a measure of relevance for attributes within singular concepts.

### 3.3 Cognitive Theories of Object Identity

We have already pointed out that nearly all research on concepts in cognitive psychology is research on categories of objects and less effort has been made studying how people represent unique individuals and how these representations (i.e. singular concepts) support our ability to identify these entities and trace their identity across time.

The study of singular concepts have been for long time neglected in favor of the study of general concepts and even when the researcher's attention has been called to the importance of individual concepts, the latter have been considered or as auxiliary to the representation of general concepts or less relevant to understand the identification mechanisms compared to higher level conceptual representations.

Exemplar models of categorization [159, 129, 172], for instance, support this difference of status by proposing that people represent categories by means of stored exemplar information. According to this view, for example, people represent the category of dogs as a set of individual dogs that they have stored in memory. However these models do not consider exemplar representations as representations whose properties are worth exploring in their own right.

The subordinate role of singular representations compared to general concepts is also assumed by a doctrine, developed in philosophy and more recently imported in psychological research, about the problem of object identity. This doctrine, known as sortalism, argues that the concept of an individual depends so tightly on the concepts of its categories that the individual's persistence, identity and distinctness derive from these concepts.

Contrary to this view, in the last years we have assisted to the development of other theories of object identity which are not based on strong assumptions about the relation between singular and general concepts. We refer to these theories as non-sortalist approaches to object identity.

In this section we look at these two kinds of approaches of object identity (i.e. sortalist vs. non-sortalist approaches) and we discuss these approaches in the light of the singular cognition problem.

#### 3.3.1 Sortalist Approaches to the Problem of Singular Cognition

Since Frege [71] first observed that one cannot count without specifying what to count, various philosophers and psychologists of language have argued that certain concepts dubbed "sortals", such as "dog", "table", "person" provide principles of individuation and numerical identity. These concepts tell us what

to count as one instance of something and whether something is the same one as what we have seen before [258, 107, 142].

Sortals are lexicalized as count nouns in natural languages that make the count/mass distinction. A sortal term, such as *table*, allows us to single out individual tables so that we can count them. On the contrary, a predicate like *wooden* denotes a property that does not by itself aid in singling out and enumerating objects. We cannot count the wooden stuff that composes a table. Should the wooden table be counted as one or should the top and the four legs be counted separately so that we have five wooden stuffs? In contrast, a request to “count the wooden tables in this room’ will receive a definite answer: a wooden table (with its legs and top) should be counted as one table. Hence the count noun “table” gives us the principles for what to count as one table, whereas the adjective “wooden” does not provide principles of counting. In general, other predicates besides count nouns, e.g. verbs or adjectives, do not serve the logical function of providing principles of individuation. We cannot count “sleeping” or “green” unless we mean, e.g. count the naps you took or the green trees.

Sortals also provide principles of numerical identity. We cannot ask the question “is this the same  $X$ ” without using a sortal to specify what  $X$  is.

When a person dies, even though we can trace a spatiotemporally connected path from the person to its body, we nevertheless decide that the person has gone out of existence. Your child and your sister’s child are two different children, whereas a certain baby and a certain grown teenager may be the same person.

This is because an individual can undergo a variety of changes in its properties, but some changes are non compatible with its identity. The distinction between possible and impossible changes for an individual determines, at least in part, the identity of that individual. An individual  $x_0$  cannot be the same individual as  $x_1$ , if the change which could explain that  $x_0$  has been converted into  $x_1$  is not compatible with the  $x_0$ ’s type. Which changes are compatible and which are not varies across types of objects. I can totally disassembly a table and then reassembly it after a while and still say that it is the same table. In contrast, total disassembly and reassembly is incompatible with a person. Therefore, which changes are possible and which are not depends from what ultimately an object is. Following Wiggins [258], a sortal is exactly what provides an answer to the Aristotelian “what is it?” The expected answer will mention the kind to which the individual belongs, enabling one to make judgments about its numerical identity over time.

Again, adjectives and other grammatical categories do not provide such principles of identity. For instance, the question whether something is “the same wooden stuff does not have a definite answer unless we mean “the same wooden table’. In this case the count noun such “table” provide the principles of identity.

Another aspect about sortals is that principles of individuation and identity provided by sortals may override our basic criteria of identity based on spatiotemporal continuity. To borrow an example from Hirsch [107]: A car consigned to a crusher follows a spatiotemporally continuous path in the crushing process and it gradually becomes a pile of metal and plastic, but nevertheless at some point, we decide that the car has gone out of existence. This is because the sortal car provides the criteria for what counts as a car.

Wiggins [258] have proposed a further distinction between *substance* and *stage* (or phase) sortals. In a nutshell, a count noun is a substance sortal if instances of the sortal it denotes cease to exist when they cease to be members of the sortal, e.g. person, dog, tree, car. In other words, substance sortals satisfy the condition that once something is no longer an X, it is also “no longer”. For example, when a person dies, he ceases to be a member of the sortal person and he goes out of existence. Hence “person” is a substance sortal. Substance sortals contrast with phase sortals such as baby or caterpillar, which do not have this property - a baby does not cease to exist when she grows up even though she or he is no longer a member of the sortal baby. Similarly, a caterpillar does not cease to exist when it becomes a butterfly although it is no longer a member of the sortal chrysalis. For Wiggins [258], only substance sortals stand for genuine kinds in a metaphysical sense.

More extreme sortalists argue that there are no individuals at all, apart from the sortal concepts that single out them and establish their beginning and endings [61]. This assumption led some authors to formulate a sort of “logic of sortals” whose main tenet is that there are no “bare particulars”. The idea is that we cannot enumerate or trace identity without the support of a sortal. “Bare particulars” are the alleged individuals that have no properties of their own whatsoever but still serve as entities on which to hang properties.

Suppose someone is pointing at some part of the visual scene and uttering the word “that”. The demonstrative “that” may refer to a bare particular. It does not pick out an individual for which we can trace identity over time. We may be able to figure out that the person intends to pick out part of the visual scene with a table present, but we would not know whether the person is pointing to the table, a colour patch of the table, the millions of molecules of the table, or the table plus the dish that is sitting on it. The main tenet of the logic of sortals denies that we have conceptual access to bare particulars.

Sortalists claim that objects are always top-down product of their categories and suggest that we can not represent, identify and track individuals without the support of sortal concepts which provide principles of persistence and identity.

## Sortalist approaches in psychological theories

The doctrine of sortalism have been recently imported in psychological studies in order to explore whether adult's or children's intuitions about individual identity match those of sortalism. Do people think that an individual's identity depends on the sortal category to which it belongs? As pointed by Rips in [23] this psychological version of the doctrine of sortalism (or psychosortalism, e.g. "the doctrine that people think that identity and differences of individuals crucially depend on sortals") deals with people's beliefs about individuation and identity and therefore it is far from the metaphysical questions that inspire the philosophical sortalism (e.g. What is an individual object?).

In this section we focus on psychological studies which provide evidences in favor of the psychosortalism.

A first consistent attempt to show how sortals play a role in explaining identity over time comes from studies on object individuation in developmental psychology.

Object individuation is the process by which we establish the number of distinct objects in an event. When an object is seen at time  $t_1$  and an object is seen at time  $t_2$ , the question arises as to whether the same object is seen on two different occasions or whether two distinct objects are present. Three main sources of information have been proposed to be involved in this process: 1) *spatiotemporal information*, 2) *property (featural) information*, 3) and *sortal information* [265, 107].

As pointed by Fei Xu [265], *spatiotemporal criteria* include the following assumptions: 1) one and the same object cannot be at two places at the same time ; 2) no two objects can occupy the same space at the same time; 3) objects travel on spatiotemporally continuous paths. No object can travel from point  $x_1$  to point  $x_2$  without traversing a continuous path in between; in presence of spatiotemporal discontinuity people judge that there must be two objects involved. For example, the wooden table that is in front of you now is not likely to be the same object that was seen in a faraway place 10 minutes ago, because no object can traverse a spatiotemporally connected path between these two locations in such a short amount of time. These generalizations are true for all objects, regardless of their kind. *Property information* include the following assumptions: 1) if we see an object belonging to a certain kind (e.g. a dog) at a time  $t_1$  and an object belonging to a different kind (e.g. a table) at a time  $t_2$ , we infer there are two numerically distinct entities; (2) upon seeing a member of a kind at a time  $t_1$  (e.g. a red car) and a member of the same kind with a different property (e.g. a blue car) at a time  $t_2$ , we likely infer that there are two numerically distinct entities. For example, the blue car that you see now is

not likely to be the same object as the red car that you saw 10 minutes ago. Of course if you see the blue car 2 years later, the interpretation that the blue car is the same as the red car is more plausible.

The property criteria are kind-relative. Certain property changes signal a change in identity only within certain kinds of objects. For example, if you see a small screen on the table now and a big screen on the same table later, you infer that there are two numerically distinct screens. But if you see a small plant in the garden now and a larger one there a few months later, it is not necessarily the case that there are two distinct plants.

*Sortal information* includes generalizations such as 1) objects do not change kind membership; 2) if an object seen at time  $t_1$  falls under one sortal concept and an object seen at time  $t_2$  falls under another sortal concept, then they must be two objects. For example, the blue car that you see now cannot be the same object as the blue table you saw 10 minutes ago. Furthermore, property information is sortal-specific such that property differences are weighted differently depending on the kind of object under consideration. For example, if a small green plant is replaced by a large leafy one in a month, it might well be the same individual that has grown over time. By contrast, if a small green car is replaced by a large green car, it is very unlikely that they are one and the same car.

A lot of psychological investigations have focused on what criteria are employed by infants for individuating objects and deciding whether something is the same one as seen before.

In the last decade, a methodology, named violation-of-expectancy looking-time paradigm, has been developed to study the cognitive capacities of pre-verbal infants [229]. In this method, infants are shown the same event repeatedly and their looking times recorded. With each repetition their looking times decline, that is, infants “habituate”. When infants reach a pre-set habituation criterion, they are shown two displays alternately, one consistent with adults’ understanding of the event and the other inconsistent. If the infants have the same understanding of the habituation event as adults, they should look longer at the inconsistent display as opposed to the consistent one. In a seminal study, Spelke et al. [230] showed that 4-month-old infants take evidence of spatiotemporal discontinuity as evidence for numerically distinct objects. In this experiment, two screens were lowered onto the stage with some space in between them. The infant saw that a rod appeared from behind one screen, say the left one, moved to the left end of the stage, then returned behind the left screen. No object appeared between the two screens. After short pause, a physically identical rod appeared from behind the right screen, moved to the right end of the stage, then returned behind the right screen. This event was

repeated until the infant reached a habituation criterion, which was defined as the average looking time of the last three habituation trials being half of the first three trials or less. The screens were then removed to reveal one of two outcomes: the expected outcome of two identical rods or the unexpected outcome of just a single rod. The infants looked longer at the one-rod outcome, suggesting that they, like adults, had expected two rods and were surprised to see just one. When the rod did appear in the space between the two screens, on the other hand, the infants looked about equally at the one-rod and two-rod outcomes, as if undecided as to how many rods were behind the screens.

Baillargeon et al. [8] presented evidence that 5-month-old infants understand that two objects cannot be at the same place at the same time and that one object cannot be at two places at the same time. In one of the experiments, infants were habituated to a tall rabbit going behind a screen and appearing on the other side. Then the middle section of the top half of the screen was removed so that the tall rabbit should appear in this “window”. If the rabbit did not appear in the window, the infants looked longer than if the rabbit did appear in the window. But if the infants were shown two identical tall rabbits simultaneously, one on each side of the screen, they did not look longer when no rabbit appeared in the window. Infants could only succeed if they interpreted the two identical-looking rabbits as two distinct rabbits using the location information. In other words, if shown two objects simultaneously, the infants set up representations of two numerically distinct objects that allowed them to resolve an apparent violation of spatiotemporal continuity.

The results of these studies are compatible with two different interpretations of how sortal concepts can underlie the identity judgments, one assuming that infants represent specific sortal concepts such as rabbit or ball, and the other assuming that the sortal concept underlying the behaviour is physical object. Xu and Carey [269] devised further experiments to address this question.

The aim of these experiments was to produce evidence in favor of the *Object-first Hypothesis* (OFH). According to this hypothesis there are two hierarchical levels of sortals in the adult conceptual system. A most general sortal named *object* for which spatiotemporal properties provide the criteria of individuation and identity and more specific sortals rely on additional types of properties to provide these criteria. The OFH claims that infants may have the sortal object before they have other sortals more specific than object. Starting from the results of Spelke et al. [230] about the ability of four-months infants of using spatiotemporal information to trace the object identity through time, the authors conducted 5 studies using the habituation paradigm to support this hypothesis.

The first experiment replicated the results of Spelke et al. [230], showing



that 10-month-old infants use spatiotemporal information to establish how many individuals are involved in an event and to track identity of those individuals over time. In this experiment the split screen procedure was used to contrast two conditions: a *discontinuous condition* where no object ever appeared in the space between the screen and a *continuous condition* where an object traced a path continuously back and forth behind the screens, appearing in the middle. In both conditions babies looked longer at what for adults would be the unexpected outcome (respectively one object and two objects).

The remaining experiments addressed directly the OFH using a variant of the Spelke's procedure. They show babies events in which one object emerges from one side of a screen (e.g. a ball) alternating with a different object (e.g. a bottle) emerging from the other side of the same screen. Adults infer that there must be at least two objects behind the screen, referring to specific sortal membership or property information. Then the experimenter removes the screen, revealing either one object (unexpected) or two (expected). The issue was to test whether babies would make the same inference of adults, looking longer at the unexpected outcome.

If the infant is able to use the property/kind (or sortal) difference between the ball and the bottle to infer two distinct objects, she should look longer at the one object outcome. Surprisingly, these 10-month-old infants failed to look longer at the unexpected one-object outcome suggesting that they do not represent sortals ball or bottle. In a variant of the experiment which varied the sequences of habituation the authors showed that the results cannot be explained with the infants' incapacity to code properties. Xu and Carey concluded that even though the infants had encoded the properties of the objects, they did not use these differences to infer that there were two distinct object. More likely they have represent the event as an object (with ball properties) and as an object (with bottle properties). As the only sortal infants represent is physical object, then they can only use spatiotemporal criteria to individuate objects but because the spatiotemporal information is ambiguous (one screen) the infants were agnostic as to how many objects were behind the screen. Further experiments in Xu and Carey [269] showed that 12-month-old infants succeed at these tasks. This is at least suggestive that the older infants may have sortal concepts such as ball and bottle. As the two exemplars belong to two different sortals/kinds, they must be two distinct objects. The experimental evidence reviewed above suggests that physical object may be the first sortal concept infants represent and that it is not until 10 to 12 months of age that they represent more specific sortals such as ball or bottle.

Van De Walle [255] et al. extended these findings by using a manual search task. Infants saw 1 or 2 objects placed inside an opaque box, into which they

could reach but not see. Across conditions, the information specifying the two objects differed. On one-object trials, a single object was shown to infants and returned to the box, which was presented for reaching. Infants invariably retrieved the object, which was then taken away from them, after which reaching was coded. Because the object has been retrieved, reaching should be brief and cursory. On two-object trials, infants was shown two objects, either simultaneously in the spatiotemporal condition or successively in the property/kind condition. Before the box was presented for reaching, one object was surreptitiously removed. Again, infants invariably retrieved the object still in the box, which was then taken away from them, following which reaching is coded. Now, infants who have represented two objects should reach often and persistently because they should expect to find the second, missing object.

The results of two experiments show that twelve-month-old infants individuate objects in the current task when provided with either property/kind information alone or property/ kind information paired with spatiotemporal information. They reach both more frequently and for a longer duration when a second object should be in the box than when the box should be empty. Ten-month-old infants, in contrast, individuated objects in this task only when provided with unambiguous spatiotemporal information that specifies two objects.

Despite disparate information-processing demands, this pattern converges with looking time data, suggesting a developmental change orthogonal to that of executive function.

In more recent studies has been shown that several factors allow young infants to anticipate two object correctly. If 10-month-old infants are able to inspect simultaneously both the objects before start the trial, then they look longer at the one-object scene. This evidence has been explained with reference to the OFH by arguing that younger infants have high-level sortal concepts, equivalent to “physical object” which provides the sortal information that infants use in the pre-view condition. As Xu stated [265] “ for both adults and young infants, there is nonetheless a sortal physical object, which is more general than person, car, or tree. A physical object is defined as any three-dimensional, bounded entity that moves on a spatiotemporally continuous path” (p 369). Blok at al. [24] offer different criticisms to this hypothesis. First, sortal theories in philosophy typically hold that terms like thing, object, physical object, space occupier, entity, and so on, are not sortals, despite their count noun syntax, because they do not provide identity conditions (see [259] for a discussion on this issue). First of all, Fei Xu’s definition of “physical object”, namely “bounded, coherent, three dimensional physical object that moves as a whole” definitely excludes all sorts of things that we certainly need to be able to pick out (and to

recognize again and again). Her definition excludes rooms, walls, floors, ceilings, corridors, trees, roads, ponds, hills and so on. Secondly, the concept of “physical object” does not allow us to single out individual entities and count them. Going back to the previous example, we cannot count the physical objects that constitute a table; the number could again be one (the table), five (the legs and top), six (the legs, top, and the table), and so on. Moreover, strong sortal theories argue that only a single sortal captures all the identity conditions for a particular object. The idea that both physical objects and basic-level terms like table simultaneously function as sortals is in conflict with the traditional sortal view.

Therefore, these criticisms cast a first doubt on the validity of the sortalist approach to explain the empirical evidences reported above.

Moreover, it was demonstrated that younger infants can use verbal cues to individuate objects [266]. Nine-month-old infants, for instance, were presented with the is-it-one-or-two task. When each object emerged from behind the screen, the experimenter labeled it: “Look, a duck!” or “Look, a ball!” With just a few repetitions of these labels, 9-month-old infants behaved like 12-month-olds in the test trials: they looked longer at the unexpected outcome of one object than the expected outcome of two objects. Infants also succeeded when two unfamiliar objects were presented and non-sense words were used. By contrast, they failed when both objects were labeled “a toy” or when two distinct tones, sounds or emotional expressions were provided. However, it is noted that non-sense words can not provide criteria of identity because they are meaningless. Therefore, as pointed by Block et al. [24], the results in [266] could be explained without referring to the notion of sortal, but simply assuming that contrastive labels encourage infants to expect the presence of two objects instead of one.

Another study tested whether labeling alone could guide the process of establishing representations of distinct objects. Using a manual search method, 12-month-old infants were shown to be able to apply the presence of labels to determine how many objects were in a box whose content was invisible to them [267]: when infants heard the content of the box labeled with two different words, they expected to find two objects inside; when they heard just one word repeated, they expected to find only one object inside the box.

These studies converge with the results of the object individuation studies: infants expect count nouns to map onto kinds of objects at the beginning of word learning, and this expectation leads them to use labeling as a source of evidence in identifying kinds in their environment. The labeling event “Look, a rabbit!” informs the infant that she should set up a mental symbol that represents a sortal concept; the sortal concept RABBIT maps onto the kind rabbit in the world. If an object seen at a different time is labeled with a different count

noun, “Look, a dog!”, a mental symbol is then created to represent the sortal concept DOG. These sortal concepts provide the basic criteria for individuation and identity: an object that falls under the sortal RABBIT cannot be the same object as one that falls under the sortal DOG. In this sense, the acquisition of basic-level sortal concepts depends on acquiring basic-level count nouns.

Sortalists developmental psychologists, such as Carey and Xu, used these evidences to support the hypothesis that sortal concepts, such as “ball” and “rabbit” allow older infants and adults to perform correctly in a is-it-one-or-two task.

According to this view, the representational system underlying object individuation (and adopted in the is-it-one-or-two-task) is fully conceptual, drawing on kind information for decisions about individuation and numerical identity. This system is completely different from that described in 3.1, that is a mid-level vision system that establishes object file representations, and that indexes attended objects and tracks them through time. Object file representations do not depend upon categorizing individuals into antecedently represented object kinds. To a large extent, the mechanisms that index and track objects through time work the same way whether the objects are instances of familiar kinds or not and are thus mid-level in not requiring placement into conceptual categories.

The idea that the object representations of young infants are identical to those that are served up by mid-level object-based attention has been recently suggested by many authors [135, 211] and have challenged the interpretation of the empirical results discussed above as evidence in favor of the sortalist view. We have already described in 3.1.1 the object index theory. Here, we briefly review how some of recent studies on object individuation have been reinterpreted from the point of view of indexing theory.

In their seminal study on object persistence which we have describe above, Spelke et al. [230] found that young infants’ individuation judgments were influenced by spatiotemporal continuity. These findings have been interpreted by Leslie et al. [135] as a form of indexing by location. Take Spelke’s first condition: as the first object appears, it is assigned an index. The index sticks to the object as it moves along, disappearing and reappearing from behind each of the screens in turn. A single object attracts a single index. The test phase, in which the single object is seen again, concurs with indexed expectations and has little novelty. The two-object test requires the infant to assign a new index (by location) to the second object and consequently attracts additional attention. In the discontinuous condition, the first appearance of the object attracts an index that, again, sticks as the object disappears behind the screen. But in this condition the object does not reappear. Instead, another object appears from behind the second screen. Because the first index still points behind the

first screen and has not traversed the gap, a new index must be assigned to the second object. Now the infant has two indexes active, which translates into an expectation for two objects. When a single object is shown in the test, the infant has two indexes active with only one of them pointing at something. The infant looks for the “vanished” object.

Another example is the study of Xu and Carey [269]. In this study infants were shown with pairs of objects by removing and replacing them from behind a screen. The objects were placed in the infants’ view either two at a time (spatial condition) or alternating, one at a time (temporal condition). The objects always differed by kind (e.g. a shoe and a cup). Following familiarization, the screen was removed revealing either only one of the objects previously shown, or both objects. In the spatial condition, 10-month-old infants looked longer when the screen revealed a single object. However, in the temporal condition, the infants looked equally at the revelation of one and two objects. They appeared unable to infer that a shoe and a cup must be distinct objects, unless they saw both objects together at the same time. When shown the cup and shoe at different times, they did not infer the presence of two distinct objects. Slightly older infants, at 12 months, successfully inferred two objects under both conditions. Apparently, Xu and Carey’s younger infants individuated only by location. Indexing theory, drawing on independently motivated notions, provides a ready explanation. Because objects are indexed by location, seeing two objects in different locations at the same time forces the assignment of two indexes: therefore two individual objects are inferred. However, the index does not automatically carry featural information. When only one object at a time is in view, only one index is assigned. The featural differences across successive appearances might be registered and remembered in the infant’s feature map, but they do not force the assignment of distinct indexes. Under these conditions at 10 months, indexing appears to be driven by the “where” and not by the “what” system. By 12 months, however, the featural differences across time apparently do force assignment of a second index. One intriguing hypothesis is that the change between 10 and 12 months in these tasks reflects increased integration of ventral (“what”) and dorsal (“where”) neural systems.

This interpretation of the results challenges the sortalist view of object individuation and suggests that the heart of any object representation might be inherently abstract, a kind of mental pointing at a “this” or at a “that”. This idea of a deictic system at the basis of object representation is in line with our notion of singular concept in that it creates the first route for a direct connection between an object in the world and its representation in the cognitive system.

Other evidences against the sortalist view have been provided by studies that have explicitly investigated the role of sortal concepts in identifying objects. We

review these studies in the next section.

### 3.3.2 Non-sortalist Approaches to the Problem of Singular Cognition

If the main tenet of the sortalism is that sortal concepts, such as “car” or “table” give people the means to identify objects, one way to provide evidence in support to or against the sortalist view is to examine the role of sortal terms in identifying objects.

As pointed by Rips in [23], a first attempt in this respect has been made by Liittschwager [137] which used a transformation method to examine children’s and adults’ willingness to attribute the same name to people after a transformation. Participants (4-year-old children) were presented with illustrated stories about people who were described as magically transformed to different states. The transformations ranged, across trials, from minimal changes of temporary properties (a clean child to a dirty child), to more extreme trans-category changes (a child to a rabbit). For each type of transformation, participants had to decide whether the transformed object could still be called by the name of the original person (e.g., Do you think that now this is Ali?). According to sortal-based theories, objects cannot maintain their identity across changes in sortal categories; so participants should use the same proper name only if the transformation is within the basic-level category person. The results of the study showed that adults as well as children were less willing to attribute the original name to the final product of the transformation the greater the transformation distance between them. The interesting result of the study was that there was no clear breakpoint on this continuum - in particular no elbow was found where the transformation crossed the sortal category boundary. These findings provided a first evidence in favor of the hypothesis that identity judgments can be maintained across changes up to the sortal category.

Using the same transformation method, Sergey Blok, George Newman, and Rips [23] reported other findings in contrast with the sortalist view. Participants read stories about an accountant (e.g., Jim) who was the victim of a serious traffic accident. As a result, Jim’s brain was transplanted in a new body: either a robot body or a human body. In both cases the Jim’s original body was destroyed. Participants had to decide whether the result of the operation was still Jim and also whether he was still a person. To investigate other factors which may contribute to judgments of identity continuity the authors introduced a further manipulation. Some participants were told about a brain transplant, while others were told that memories from the original brain were copied onto a computer, placed in control of a robot or humanoid body. The

most important prediction of the study was the following: if people use sortals to guide identity judgments, they should be no interaction between whether Jim’s brain is transplanted into a human or robot body and whether the question of continuity is about being a person (category judgment) or about being Jim (identity judgment).

Contrary to this prediction, the results indicated a dissociation between identity and category judgments. In determining whether the creature post-operation was still Jim, participants paid more attention to whether Jim’s memories were preserved and paid less attention to whether the recipient of these memories was a robot or a human body. Therefore, in some conditions participants were more likely to agree that the creature was still Jim than it was still a person and vice versa in other conditions they were more likely to agree that he was still a person than it was still Jim. This double dissociation presents difficulties with the sortalist view, since this theory predicts that Jim’s existence should cease when he stops being a person. On the contrary the results indicated that Jim continues to exist, though out of his sort, that is out of person category.

In a second transformation study the authors tested whether natural kinds and individual artifacts can persist across changes in sortal categories. Participants were presented with a picture and a short description of an object (e.g. a particular cat). They also were shown with a drawing of a sci-fi device which could be a “transporter” (i.e. a device to transport the object particle-by-particle to a new place and reassemble it) or a “copier” (i.e. a device that made a new copy of the object, while the original was destroyed). Finally, participants saw a picture of the outcome of the transformation that could be the same picture as before (e.g. the picture of the same cat), a picture of a related object (e.g. a picture of a dog) or a picture of an unrelated object (e.g. a picture of a boat). The task was to judge whether the outcome of the transformation was still the same individual and whether it was still a member of the same sortal (e.g. a cat).

As in the previous experiment, the results produced a dissociation between the identity question and the category question. When the outcome was the same as the original, participants judged the outcome to be a member of the same category but were less convinced that it was the very same individual. On the contrary for transformation involving related objects the pattern of results was reversed. They were more likely to agree that the outcome was the same individual than that it was a member of the same category. No differences were found for transformations with unrelated objects.

Another non-sortalist approach to explain judgments of the persistence of individual objects has been recently proposed by Rips et al. [192]. In a se-

ries of evocative studies, the authors examined the role of causality in identity judgments and have proposed a new model of object identity named *Causal Continuer Theory*. The model derives from a philosophical theory, i.e. the Closest Continuer Theory, proposed by Robert Nozick [174] as a theory of personal identity. We have already noted that the problem of object (or personal) identity deals with the question of how people decide that an individual object, let's say  $x_0$ , existing at one time is identical to one of a set of candidate objects,  $x_1, x_2, \dots, x_n$ , existing at a later time. The Nozick's theory suggests that the identical object to the original  $x_0$ , i.e. the continuer, is the one that is, in some ways, the closest to it. In the Rips et al.'s model this closeness is explained in terms of causal dynamics and therefore the model has been referred as *Causal Continuer Theory*. The intuition is that "the continuer of the original object must be a causal outgrowth of that original" (p. 7). Causal continuity captures the intuition that people think of causes as central to object persistence and suggests that what makes two entities identical with each other is not based on superficial similarity or sortal membership but rather on a deep causal connectedness.

While the first element of the model deals with causality, the second element deals with closeness. As we noted above, the model assumes that in determining a continuer, people do not select something that is arbitrary far from the original. If there are two or more objects at a later time that are close enough to the original, the theory specifies that only the closest of these objects is identical to the original. However, if none of these potential continuers is significantly closer to the original than the others, the model predicts that indecision can be generated due to the competition between the candidates. Another aspect related to the closeness is that in determining a continuer, people can not to select something that is arbitrary far from the original. If the candidates are causally too far from the original there may be no object that qualifies as identical to the original.

In the causal continuer framework, the authors proposed a two-step decision process on identity judgments. 1) The first step deals with considering as potential candidates only those objects that are close enough to the original; 2) the second step consists of selecting, within the range of candidates, the closest object as the one identical to the original.

Note that the determination of the range of candidates is context dependent. This means that an item in one situation may not be the closest in another if the second situation contains an even closer object.

In a series of experiments Rips et al. [192] evaluated the psychological plausibility of the model and they developed a quantitative version of the model that provided accurate predictions about identity judgments in different tasks.



In the experiment most relevant for the present discussion about sortalist versus non sortalist approaches, the authors tested the hypothesis that causal factors rather than sortal category membership dominate judgments of object identity when these factors are contrasted. In this experiment, participants were asked to make an identity choice between potential continuers. The causal distance between the continuers and the original object was systematically varied across the trials. Participants were presented with stories about a machine that could copy and transfer objects from place to place on a particle-by-particle basis. The machine copied the particle of the original object and retransmitted them to a new location where the particles were reassembled while the original particles were destroyed. The duplicating process was interpreted in the experiment as the guaranty of the causal connection between the original and the continuers, whereas spatio-temporal and material connections were eliminated by the fact that the particles were reassembled in another place and the original particles were destroyed.

The causal distance between the original and the continuers was varied by changing the proportions of particles in the copy that the original object causally produced. In order to contrast the predictions of the causal continuer theory with those of the sortal theory, the source of the particles which completed the outcome of the transformation (when less than 100% of the original particles stemmed from the original) was varied. On half of the trials the residual particles were from another member of the original's category, whereas on the rest of the trials the residual particles came from a member of a different basic-level category. The task was to decide 1) whether the copy was the same object as the original; 2) whether the copy was in the same category as the original. Moreover in one condition (one-copy condition) a single copy was generated from the original, whereas in another condition (two-copy condition) two copies were derived from the original. Therefore, in the second condition participants were asked to choose whether only one copy of the two copies was the same object as the original (or in the same category as the original); whether both copies were the same as the original (or in the same category as the original); or finally whether neither of the copies was the original (or in the same category as the original).

Confirming the results reported in [23], the authors found a dissociation between category judgments and identity judgments. The larger the percentage of particles from the original, the more likely participants responded that the copy was the same as the original. However, no effect was found of whether the residual particles were from a member of the same category or of a member of a different category compared to the original. The opposite pattern was found for the category judgments. When more and more particles came from the opposite

category, more likely participants judged improbable that the outcome of the transformation was a member of the same category as the original. The dissociation indicated that factors affecting category membership do not necessarily affect decisions about individual persistence. This results are in contrast with the assumptions of the sortal theory which predicts that factors that cast doubt on whether the copy is a member of the sortal category should also cast doubt on whether the copy can be considered the same individual as the original, contrary to the results.

Moreover, the data from the two-copy condition supported the prediction of the Causal Continuer Model, showing that as the percentage of original particles in the two copies became more dissimilar, participants shifted toward judging that only the dominant copy (i.e. the copy with the largest number of particles stemmed from the original) was identical to the original. On the contrary, when both copies had the same number of particles from the original, neither copy was dominant and participants were more likely to judge that both copies were identical to the original or neither copy was identical to the original if the number of particles stemmed from the original was scarce. Again no difference resulting from whether the residual particles came from a member of the same category as the original or from a member of the contrast category, confirming the results of the one-copy condition and reinforcing the challenge toward the sortalist view.

Even though other authors [190, 190] tried to defend the sortalist approach, the empirical results from Rips et al.'s studies reinforce the doubts about sortal theories since they were obtained from experiments which were thought as systematic attempts to pit the causal theory against alternative theories and in primis against the sortal theory.

Moreover, the model in Rips et al. [192] aimed to capture adults' judgments about object identity over the long term, investigating what people take to be the ultimate basis for object identity even in situations in which perceptual information was not involved. On the contrary, sortal theory in psychology was originally applied to research on infants and children by Macnamara [146] and Xu and Carey [269] to investigate object individuation across temporary perceptual interruptions. Although the results of developmental studies may be relevant to adult judgments, there is no empirical evidence that directly links these two programs of research, and the lessons from the infant research are ambiguous, as we have noted above.

Beyond the debate between sortalists and non-sortalists, these results are also highly relevant to the present work because they suggest that singular representations of individuals (i.e. singular concepts) can free themselves from the bounds of basic level categories, becoming the very tokens on which people base

their identity judgments. In other words, these findings show that people do not believe that knowledge of individuals depends so tightly on knowledge of categories that their identity can not be preserved across the category boundary. On the contrary, it appears that people may have representations of individuals apart from their representations as members of a category. We argue that these representations, which we refer to as singular concepts, possess their own individuality in the conceptual system and are indeed the core of object identity. Our aim is to show that many psychological theories have overstated the dependence of individual and general concepts and many aspects of singular cognition can be explained without assuming that general concepts provide the ultimate principles to organize the knowledge about individuals, the access to this knowledge and the use of it in identity judgments.

### 3.4 Proper Names as Index of Individual Identity

When we have described the nature of singular concepts in Chapter 2, we have claimed that singular concepts may contain a special kind of information, corresponding to the proper name of the entity represented. We argued that these “mental proper names” serve as a sort of mental labels which can be used to create a direct referential link between the singular concept in memory and the corresponding individual in the world, even when the object is not directly perceived. Insofar as they serve as longstanding labels on singular concepts, which can be used for accessing, adding, updating, and merging of information on an individual, even without the need of a perceptual contact with that individual, mental names cannot be pure demonstratives or indexicals, which are contextually based determiners of their objects. In this sense they differ from visual indexes described in section 3.1.2, but they serve as the cognitive counterpart to the proper names that are used in language to refer to unique individuals. Indeed they are the mental encoding of the proper names used in language.

Many evidences confirm that proper names are processed differently within the cognitive system than other kind of information.

In this section we review the studies which support our assumption that proper names have a different status within the mental representations about individuals.

A first evidence comes from neuropsychological findings that show that proper names follow functionally distinct processing pathways compared to common names.<sup>2</sup> Neurological damage can result in a condition whereby only proper

---

<sup>2</sup>The common names/proper names distinction is interesting from our perspective because

names are disturbed, while common names are unaffected. The opposite condition whereby proper names are spared but there are severe problems with common names has also been observed, even though the latter condition has been reported less frequently than the first condition (see [216] for a review about neuropsychological dissociations between common and proper names). Taken together, these two conditions, mirroring each other (thus constituting, what is known in neuropsychology as a “double dissociation”), constitute evidence of a separation of mechanisms processing proper and common names in the brain. An interesting interpretation of these evidences have been recently proposed by Semenza [216] within a theoretical, information-processing model of proper name production and understanding which is based on the notion that different ways of possessing reference, which distinguish proper names from common names, are reflected in different mechanisms governing semantic memory, distinguishing “individual semantics” from more general semantics. In particular, the properties of semantic operations necessary for naming with proper names may have, in comparison with those used for common names, different qualities. As Semenza, Zettin and Borgo [217] have observed, a name designating a category applies to a set of attributes overlapping or interacting with each other via high-probability connections. The set of attributes labeled by a proper name, instead, combine together incidentally, being related to each other only by virtue of belonging to entities that are unique. This fact would explain why the link proper names have with their reference is more likely damaged than the link that common names have with the objects they label.

Another evidence which confirms that a proper name has a different status within a singular representation, derives from studies showing that proper names are more difficult to retrieve than is biographical information about people. We have already discussed these studies in 3.2 showing how they influenced a number of theoretical accounts of person recognition and naming.

Other studies have investigated the cognitive relevance of proper names in the context of developmental psychology. These studies have been reported by Jeshion in [115] as evidences supporting her view about the significance of proper names. We review here some of these studies that show the special status of proper names in the cognitive reference system.

Several studies of word learning in childhood have explored the question of how children learn proper names.

By the time they are two years old, children appear to know which expressions in their language are proper names, and they also genuinely seem to

---

reflects the distinction, at a conceptual level, between singular concepts and general concepts. Proper names essentially refer to individuals (or individual groups), while common names refer to categories. In the same way, singular concepts represent individuals, whereas general concepts represent categories.

represent these expressions as designating individual objects. Children’s ability to use syntactic and semantic clues to comprehend new names was investigated by Katz, Baker and Macnamara [125] and Macnamara [145].

In these studies, the authors showed one group of 2-year-olds a target doll and labeled it with a novel word modeled syntactically as a proper name (e.g., “This is ZAV”). These children restricted the word to the target object and were unwilling to extend it to another similar-looking doll. This finding suggests they interpreted “ZAV” as a proper name designating the individual target doll. In contrast, another group of 2-year-olds heard the same target doll labeled with the same word modeled syntactically as a count noun (e.g., “This is a ZAV”). Children in this group readily extended the word to both the target and another similar-looking doll, suggesting that they interpreted the word as a count noun picking out an object category.

But how do children acquire the ability to use syntactic information to identify proper names in speech? In her paper on significance of proper names Jeshion [115] [p. 383] noted that “knowledge of syntax helps guide children in identifying novel words as proper names as opposed to common nouns” , but “while common nouns taking determiners provides a basis for distinguishing them from proper names, syntax is insufficient for proper name identification”. The author pointed out that other possible candidates, such as pronouns, adjectives and mass nouns, could be considered to identify the novel words modeled syntactically as proper names (e.g. “this is Zav”). And still remains open the question of how a child is able to acquire syntactic knowledge. According to Jeshion, this would imply assuming that the child “must be able to already possess some knowledge of the linguistic category of some words”.

One proposal is that children use semantic information (i.e. knowledge about the properties of real-world entities) to learn these words’ syntactic markings.

According to Macnamara, for example, children rely on the assumption that proper names are the words that people use to pick out objects belonging to kinds of things whose members are seen as enduringly significant in their own right.

One such assumption is that only some kinds of individuals are regarded as candidates for a proper name. There are empirical evidences that support this hypothesis. Gelman and Tylor [83], for example, showed that children in their experiments exhibited a strong tendency to choose an animal-like toy as the referent of a proper noun but they were reluctant to assign proper names to artifacts-blocks, shoes, toy cars and planes. Similarly, Hall [99] showed that 3- and 4-year olds made proper name interpretations when they learned novel words for typical pets, like birds or dogs, as they did the majority of those who learned novel words for nonstandard pets described as possessed by the

experimenter.

Another evidence supporting the hypothesis that children, like adults, represent proper names as referring to unique individuals (i.e. Unique Individuals hypothesis) has been provided by a study of Sorrentino [228].

Using a very simple experimental procedure, children and adults were shown an object (a toy animal or a non-animal artifact) with a salient property (an object marker like a colorful bib). The object was introduced with a novel word (“This is daxy”). The object was then moved, the object marker was removed and a second object identical to the first was introduced at the location before occupied by the first object. The object marker was placed on the new object. At this point the experimenter asked “Which one is daxy?”. In the animal condition both children and adults selected the toy animal originally referred to with the new word (despite a change in the animal’s appearance and location). Note that this new word was interpreted by adults like a proper name in a evaluation session that followed the main task. The results show an interesting bias named “animal bias”: participants in the artifact condition did not interpret the word as a proper name.

This study provides evidence in support of the central role that proper names plays in reference. Proper names are paradigmatic referring expressions not only from a philosophical point of view but also from a psychological point of view. The results support that proper names contrast referentially with other referential expression like count nouns, mass nouns or adjectives. Proper names are represented as referring to unique individuals, or rigid designators, namely they refer to unique individuals and are used to trace the identity of their referents through changes in appearance and location. Another interesting point is about *the animal bias*: children like adults readily learn proper names for people, many animals and their surrogates (e.g. dolls) but not for non-animal individuals such as wooden blocks. In this study adults interpreted the new word introduced in an ambiguous sentence frame as a proper name when the referent was an animal but not when the referent was a simple artifact. This result is interesting because it shows that there are type of entities that are conceived more prototypical namable entities. In this respect artifacts represent a special type of entities, an aspect which we have investigated in our work. People may assign proper names to complex artifacts such as boats or cars (e.g. Titanic) but generally they do not assign proper names to simple artifacts (like bottles or knives). It would be of interest to investigate the properties of objects that lead people to consider them like candidates for proper names. If proper names are means to trace personal identity, we can suppose that people assign proper names to objects that are cognitively not interchangeable with other objects. Being good candidates for proper names could depend more from a

differentiation need than from category membership.

Some interesting insights about this issue have been provided by Hall et al. [98]. In two studies the authors explored 5-year-olds and adults beliefs about entities that receive reference by proper names. This research was motivated by two main goals: 1) to develop a set of norms about the structure and coherence of children's and adults' concept of proper namable entity, 2) to develop a possible account of how children learn proper names. To this purpose, authors adopted two different tasks. A listing task in which children and adults stated what things in the world can or cannot receive a proper name and an explanation task in which participants explained why things receive proper names. The results show that children tended to list animate living things and their surrogates as meriting a proper name and tended to not include human artifacts or other things as being deserving. In contrast the lists reported by adults included all the previous type of things. Children and adults showed similar belief about non namable things, tending to report artifacts and other things as being unworthy of a proper name. Both groups of subjects provided similar explanation for why things can receive proper name. The main explanation is that things receive proper names in order to be identified as individuals in their own right or to be distinguished from other things. The only alternative explanation is the need to interact with something socially or to mark affection for it.

The results of this study are interesting at least for two reasons. First, they show that the difference between children's and adults' beliefs about entities that receive proper names does not depend from a different explanation of why certain things are deserving of reference by proper names but it seems more related to a different experience and knowledge of things that receive proper names. This result is interesting because it provides evidence of a common referential mechanism for children and adults that goes through a gradual specialization in the course of the development to include more types of namable things. The second important result concerns the non-namable things. Both children's and adults' lists of non-namable objects were heavily filled with artifacts and lesser degree with other things. This result confirms the peculiarity of artifacts (and their difference respect to living things) also in reference. Moreover the fact that some artifacts were considered as deserving of proper names by some participants and as not deserving of proper names by others show that there can be difference in terms of how people construed the items. We claim that a possible difference could be derived by a not clear differentiation between proper names and brand names. In the present work we will investigate another kind of label for artifacts, i.e. the model name, that often includes or is strongly associated to, the brand name (e.g. Fiat 500). As we will discuss later, our assumption is that model names can function as proper names to label a special

kind of individual concepts, that is singular concepts of products. This aspect has been fully investigated in the experiments reported in Chapters 6 and 7.

The thrust of this section was twofold. First, we aimed to describe empirical evidences supporting the assumption that proper names have a different status in entity-specific information processing. Second, we wanted to provide evidence in favor of the idea that there is a connection between names and significance. This connection recalls the principle of significance discussed in Chapter 2 for the initiation of singular concepts. This suggests that individuals that are deserving of reference by proper names are more likely to be the individuals for which we have singular concepts in memory, because both mechanisms, i.e. conferring proper names and initiating singular concepts, have the same aim to signal the individuality of individuals.



## Chapter 4

# Neural Basis of Singular Concepts

The way in which we interact with the world is determined by our network of accumulated knowledge concerning the people, objects, places, animals and all the other entities that comprise it. This knowledge is stored in a sort of mental repository typically called semantic memory. Within of semantic memory we have distinguished between knowledge about general categories (general concepts) and knowledge about individual entities (singular concepts). The focus of our work is on the access, the functioning and the organization of this conceptual database in its part concerning the representation of individual entities.

In the chapter 3 we have discussed cognitive models which assume that the cognitive system uses singular representations of individuals and mechanisms for direct reference to objects in order to support processes concerning the tracking (perceptual or conceptual) of unique entities and their identification. We have also contrast these models with other approaches (sortal theories) which suggest that singular representations are in subordinate position compared to high level representations (sortal concepts) to guarantee object individuation and persistence across time.

In this chapter we discuss the studies which provide evidence for the existence of brain areas involved in representation and processing of singular representations and we review the literature that have investigated different aspects concerning with the neural basis of singular concepts and singular cognition.

## 4.1 Neuropsychological Evidences

Neuropsychological studies of brain damaged individuals with disrupted semantic memory, and more recently, functional imaging investigations in healthy control subjects, have been critical in shaping current theories regarding the cognitive and neural architecture of semantic memory. Many studies have demonstrated that different parts of brain may be selectively involved in processing of different categories. Subjects with deficits affecting one category of knowledge, with relative preservation of another, have been of particular interest in this respect [106, 256, 117]. In particular, different systems appear to mediate the access to knowledge about tool and man-made objects as opposed to natural categories of objects [40, 126]. Such findings are accepted by some researchers as reflecting categorical organization of semantic memory, with separate representation of distinct domains of knowledge [40].

Apart from representing categorical information, the human brain continually deals with a vast amount of specific knowledge about individual entities. However, the characterization of entity-specific semantic knowledge as a dissociable domain from general knowledge concerning general categories has been less investigated.

The majority of studies that has addressed this issue have investigated the neural basis for processing knowledge about individual entities within a specific category (e.g. faces), focusing on a specific process such as visual recognition, access to (specific) semantic knowledge or naming.

However, very few studies have compared entities from different categories which can likewise be accessed at the exemplar level, matching the level of categorization at which the stimulus is processed (e.g. comparing a famous building like the Eiffel Tower with a famous person like Marilyn Monroe).

In this section we aim to show how when this match is ensured, the evidence for a category specialization appears less strong, showing that many areas activated for processing individual entities of a category are also activated for that of entities of another category.

One of the most studied category of unique entities is that of familiar people. This is not surprising if you consider that the ability to recognize and distinguish a person from another is a fundamental skill necessary for the everyday social interactions.

The common issue underlying these studies is to explore whether specific neurocognitive systems are involved in person-specific knowledge processing. This issue is closely associated with theories of modular specialization in the brain. In particular, the postulated existence of a brain area that is specialized for face processing would provide a clear example of domain specificity, one of

the defining features of cognitive modules [70].

Two sources of evidence are available regarding the functional anatomy of the different stages of person identification: the association of deficits with the lesion sites and functional imagining.

Neuropsychological studies described patients with impairments at various stages of the person identification process, including 1) a presemantic stage when recognition of famous faces is impaired only in the visual domain (i.e. prosopagnosia); 2) the semantic stage when loss of biographical information about known people occurs regardless of stimulus modality (crossmodal agnosia for familiar people); and 3) the post semantic lexical retrieval stage, when name retrieval is impaired but semantic information is retrieved correctly (i.e. proper name anomia).

The issue whether these deficits reflect the existence of face or person-specific cognitive modules has been debated since the earliest reports of prosopagnosia.

Some studies provided evidence for the selectivity of the disorder for faces as opposed to other types of objects. De Renzi [188, 189], for example, described a severely prosopagnosic patient who easily performed a variety of subtle visual recognition tasks with objects such as wallets, neckties and photographs of cats.

The patient studied by McNeil et al. [154] presented a very severe prosopagnosia with a stable and longstanding impairment for recognizing very familiar people. Nevertheless, he was able to identify another group of visually and easily confusable stimuli, the faces of sheep.

The idea that prosopagnosia is an impairment of a specialized form of visual recognition that is necessary for face recognition but it is not necessary for common object recognition, is suggested also by a study by Farah et al. [64]. In two experiments the authors found that when a prosopagnosic patient was asked to discriminate both faces and visually similar exemplars of non-face object categories, the patient performed disproportionately poorly with faces compared to normal subjects. The results were used to disconfirm the hypothesis that the dissociation between face and non-face recognition in cases of prosopagnosia is due to a greater difficulty of face recognition, requiring an higher level of discrimination within category, compared to object recognition.

Although they cannot recognize people from their faces, prosopagnosic patients are frequently able to identify people from their voices or clothing. They can also retrieve semantic information about these people in response to their name [56].

However, other studies have investigated the selective impairment for person-specific semantics. Hanley et al. [103], for example, described a patient who, following herpes simplex encephalitis, had difficulty in identifying people from their face, their name, and their voice and was unable to gain access to mental

representations of precise person semantic information.

A selective deficit for person-specific knowledge with relative preservation of general semantic knowledge has been reported by Thompson et al. [245]. In their study, the authors described two patients which presented contrasting patterns of semantic memory deficits (i.e. impaired person-specific semantics, with relative preservation of knowledge about objects and animals and vice versa) and lateralized temporal lobe atrophy (right vs.left).

Finally, cases were described which presented a specific impairment in retrieving proper names of familiar persons, even though the access to person specific knowledge about them were preserved.

Lucchelli and De Renzi [143] reported the details about a patient that, following a left thalamic infarct, showed a marked impairment in retrieving person proper names in response to faces and to verbal descriptions, despite being able to provide precise information about the persons he could not name and to point to their photograph when the name was provided by the examiner.

A very similar deficit is also reported by Fukatsu [72] which described a patient with a selective deficit in retrieving proper names after left temporal lobectomy. He showed proper name anomia in conversation, in response to photographs, and in verbal descriptions, despite being able to provide semantic information about the people he was unable to name.

These studies seem to confirm the existence of dedicated cognitive modules for faces. According to this view, faces are very unique stimuli and are thus served by specific dedicated systems. Face-specific deficit at different stages of the identification process are also in line with the most accepted cognitive models of person recognition which distinguish between different stages involved in the process of recognizing and accessing information about people (see for example [35, 250, 38])<sup>1</sup>.

However, although a number of cases have emerged in whom knowledge concerning familiar or famous people has been severely disrupted with relative preservation of other domains of semantic memory [62, 103], in other cases the deficit extended to impaired recognition and naming of specific objects that, like faces, have many visual similar neighbors, e.g. breed of dogs, types of flowers or cars, individual animals, familiar building and landmarks [62, 54, 53].

Particular relevant for our work are the impairments of unique familiar stimuli (e.g. landmarks) referred as “semantically unique items” by Gorno-Tempini [90], who described them as items “which carry unique semantic associations that are not shared by other perceptually similar category members” (p.2087).

Unique semantic stimuli such as landmarks and building have been often reported as impaired not only in prosopagnosia, but also in patients with person-

---

<sup>1</sup>A more detailed description of these models is reported in Section 3.2

specific deficits in semantics and lexical levels. Likewise, patients with deficits in memory for specific persons have been noted to have parallel defects in memory for these other kinds of specific entities.

Gentileschi et al. [84], for example, reported the case of a woman with no focal brain lesions, suffered from a progressive impairment in recognizing familiar people along with an impairment in recognizing of famous buildings and songs.

In a systematic review on selective disorders in recognition of familiar people, Gainotti [74] compared the recognition of unique entities (famous monuments, cities, countries and other geographical entities) with recognition of familiar people in individual and group studies of patients with right and left temporal lobe lesions. The analysis showed that the recognition of unique entities, in particular monuments, was impaired in almost all the subjects with right temporal lobe lesions. On the contrary, in patients with left temporal lobe lesions identification of unique entities was spared. However, an impairment in finding entity proper names was observed when the anterior parts of the left temporal lobe was selectively damaged.

In a similar vein, some patients with anomia for proper names of famous person have been noted to have parallel defects in naming geographical items.

For example Otsuka [176] has recently described a patient with proper name anomia following subcortical hemorrhage in the left superior temporal gyrus. Despite the preserved ability to retrieve common names, the patient could not retrieve the names of people, countries, or racehorses, which he could recognize quite well and whose semantic knowledge could be accessed.

Also the anomic patient studied by Semenza [217] presented a significant impairment for geographical items, in addition to the deficit in producing proper names of famous persons.

The impairment in accessing knowledge about unique entities belonging to different categories has been also reported in amnesic patients.

In a case study of an amnesic patient with a medial thalamic lesion, Miller et al. [162] found that their patient was unable to access information about unique entities across a range of domains (e.g. famous people, events, famous buildings, movie titles), while his memory for more general knowledge was intact.

A modality specific semantic knowledge loss for unique items is also reported by Kartsounis [124] who described a patient who had great difficulty in identifying in the visual modality historically known people, such as Queen Elizabeth I and Napoleon, and well known world and London landmarks, such as the Parthenon and Buckingham Palace.

Such co-occurring deficits seem not consistent with the existence of person-specific modules and are more consistent with the hypothesis that co-occurring

deficits for unique entities of different categories can arise from the damage of a system for stored knowledge of unique identity.

When an individual entity is identified at the unique level of identity (recognized as familiar or named with its proper name), the identification process requires that the entity is distinguished from other perceptually similar category members and it needs to be linked to unique semantic and lexical knowledge. According to this view, the process of identifying unique faces could be not consistently different from that of identifying unique entities of other classes. In this perspective, all the unique entities which have unique conceptual and lexical associations are akin to the category of known persons and might share neuroanatomical or functional underpinnings.

This view is in line with the explanation proposed by Gentileschi et al. [84] to explain the deficits of their patient. The authors argued that the patient's difficulty in identifying familiar people was the consequence of progressive loss of stored exemplars of familiar persons and also of some other "unique items" (famous songs and monuments) in an independent subsystem of semantics named "exemplar semantics".

The idea that recall and retrieval of unique items from different categories depend on a common mechanism (i.e. that is not specific for person-specific knowledge) has been proposed by Damasio [54]. He suggested that evoking unique entities depends on trigger the disparate neuronal patterns that correspond to the separate inscriptions that are associated with a unique item. Further he proposed that the rostral regions of the temporal lobes may acts as "converging zones" binding together the distributed representations of concepts and that this mechanism may mainly concerns "unique entities". When translated with our terminology, the idea of Damasio is that the anterior temporal lobes serve to trigger and synchronize feedback projections to the multiple cortical regions that hold the separate inscriptions that compose a singular concept.

In contrast, recall of non-unique entity such as identifying an item as member of a category would occur as a result of biding within and among the more posterior, single modality cortices.

## 4.2 Neuroimaging Evidences

The localization of neural circuits specialized for the identification of famous or familiar entities from different categories is an area which has recently received attention in the neuroimaging literature.

Also in this research area, many studies focused on a specific category of familiar entities and the most studied category is again the person (face) category.

It is worth to note that the comparison between familiar and unfamiliar entities is particularly interesting from our point of view, because it provides the opportunity to examine the neural systems activated when pre-existing semantic and biographical information is available for retrieval, that is when a singular concept is available in memory.

The distinction between familiar and unfamiliar stimuli is particularly important in reviewing the face processing literature, since many studies of face-specificity have focused on the perceptual level of processing (irrespective whether faces are familiar or not) [123, 186, 218, 104]. These studies have used unfamiliar faces that can not be linked to specific semantic knowledge to study initial category-specific processing (face vs. other object categories) in the extrastriate cortex, while minimizing semantic processing [123, 186].

It was found that viewing and matching unfamiliar faces relative to other categories of objects consistently activate a region of the lateral fusiform gyrus bilaterally, but more consistently on the right, that has been labeled as “fusiform face area” (FFA).

However, the face specificity of this area has been challenged by other studies that found that the response in the FFA is not exclusive to faces. FFA also responds to animal faces [147] and is activated when visually similar objects are categorized at the subordinate level, e.g. when distinguishing different types of birds or cars, especially when the subject is an expert [79] and also when expertise with novel objects (greebles) is acquired [81].

Beyond asking what stimuli an area prefers, other studies focused on asking what type of computations are performed and to what behaviors such computations contribute. In particular in the case of face processing, the issue is to understand whether the FFA is specialized for face detection (i.e. detecting exemplars from face category so that specialized routines can be engaged) or it is involved in processing faces at the individual level.

One of the first studies that addressed this issue was conducted by Gauthier et al. [82]. Exploiting the mechanism of habituation, the authors measured the activity of three different areas (two selective for faces and one selective for letters) during a task in which participants attended to the location of stimuli (faces or letters). They found that activity in the face-selective areas habituated to the repeated presentation of one exemplar more than to the presentation of different exemplars of the same category, supporting the hypothesis that these areas are involved in processing stimuli at the individual level. Indeed, if these areas were involved only in face detection, no differences in activity should have been registered.

Other studies explored the effect of familiarity on the processing of human faces. Numerous neuroimaging experiments on familiar face recognition ex-

plored the activations in the face-responsive regions of the ventral extrastriate cortex, producing often inconsistent results [167, 60, 91, 203].

Nakamura et al. [167], for example, found that the right inferior temporal/fusiform gyrus responds selectively to faces but not to non-face stimuli and it is involved in the first stages of face perception; on the contrary the right temporal pole is activated during the discrimination of familiar faces from unfamiliar faces. It is worth to note that in this study the activation of the right temporal pole was not face-specific, suggesting that this region may be associated with the recognition of familiar entities regardless of the entity category.

Based on the distinction between familiar and unfamiliar entities, Leveroni et al. [136] conducted an event-related fMRI study in which they compared brain activations associated to newly learned faces, unfamiliar faces, and famous faces. They found that the recognition of famous faces produced significantly larger MR signal intensity changes over widespread areas of the prefrontal, lateral temporal and mesial temporal regions (hippocampal and parahippocampal regions), compared to recognition of recently encoded faces or unfamiliar faces seen for the first time. These brain areas have been interpreted as a common neural network for long-term retrieval of famous face.

However, Gorno-Tempini et al. [91] found a very different pattern of activations. They observed that the areas specialized for the perceptual analysis of faces were right lingual and bilateral fusiform gyri, while the areas specialized for famous faces spread from the left anterior temporal to the left temporoparietal regions. Interestingly, the same areas activated during the processing of famous faces were activated also during the processing of proper-names.

On the contrary, Dubois et al. [60] found that the main difference between familiar and unfamiliar faces involved the early visual areas (with a decreased activity for familiar faces) and a region outside the ventral extrastriate cortex, i.e. the amygdala (which was more activated for unfamiliar faces).

Beyond the differences of localization reported in different studies, what is relevant for the purposes of our investigation is that certain areas in the brain are specialized for processing faces at the unique level of identity. Related to this evidence is the issue whether these areas are specialized to identify unique exemplars from face category or they are involved in processing unique entities from different categories. To answer this question, other studies compared the patterns of activation for familiar entities from different categories (e.g. face and landmark) in different tasks.

Converging evidences from functional neuroimaging studies showed that tasks requiring processing of famous entities from different categories (e.g. persons or landmarks) activate similar neural regions [90, 167].

In particular anterior temporal regions consistently have shown stronger re-



sponses to a variety of familiar stimuli, including faces, names and landscapes.

Gorno-Tempini and Price [90] investigated in a PET study the effect of fame on activation elicited by famous and non-famous faces and buildings during a same-different matching task. They found that the task elicited constant category-specific activations in the fusiform and parahippocampal/lingual areas which was not modulated by fame. On the contrary the activation in the left anterior middle temporal gyrus showed an effect of fame that was common for faces and buildings.

The results suggested that the left anterior temporal cortex is involved in shared analysis of unique semantic attributes which mediate the identification at the unique level of identity.

Unique and non-unique entities were also compared in studies which explored specialized brain areas involved in retrieving of names for unique items.

In a PET study Grabowski et al. [92] investigated the role of the left temporal pole in naming unique entities (famous faces and landmarks). The author tested the hypothesis that cortices in the left temporal pole are engaged when lexical retrieval was performed at unique level. To this purpose they used two categories of unique entities, face and landmark, and studied the activations in the left temporal pole during a naming entity task. The PET results showed a significant activation of the left temporal polar region for both unique naming tasks (person and landmarks naming) when compared to the baseline tasks using unfamiliar entities. Interestingly, the authors found that retrieval of proper names of persons and landmarks engages the left temporal pole to a comparable degree, supporting the hypothesis that the same brain region is linked to the level of specificity of word retrieval (i.e. the retrieval of a proper name referring to an individual entity) rather than the conceptual class to which the stimulus belongs.

It is interesting reading the results of the Grabowski's study in the light of previous evidence from both lesion and functional imaging which implicated relatively segregated sectors of inferotemporal (IT) and temporal polar (TP) cortex in the process of word retrieval for entities belonging to different conceptual categories.

Tranel et al. [247], for example, tested a large sample of subjects with focal, unilateral brain lesions using a procedure which required the visual recognition of entities from three categories: unique persons, non-unique animal and non-unique tools. Results showed that defective recognition of persons was associated with maximal lesion overlap in right temporal polar region; defective recognition of animals was associated with maximal lesion overlap in right-mesial occipital/ventral temporal region and also in left mesial occipital region and defective recognition of tools was associated with maximal lesion overlap in

the occipital-temporal-parietal junction of the left hemisphere.

In our perspective, the relevant finding of the Tranel's study was that the unique-level recognition and naming of face stimuli were associated with the activation of a segregated sector in left TP and a sector of the left middle temporal gyrus but not in left ventral and posterior IT, which were engaged in recognition and naming of non-unique stimuli from animal and tool categories.

Since in the Tranel's study faces were used as unique-level stimuli and because faces are special entities for a variety of reasons [54], it is possible that the activation of anterior temporal regions reflects the specialization of these regions for face processing and not necessarily for unique level processing.

In this vein, the results of the Grabowski's study provide a very important evidence supporting the hypothesis that the effect found when naming persons is linked to the level of specificity of the retrieval, rather than to the special properties of face stimuli.

The hypothesis that the categories of famous landmarks and famous persons share neuroanatomical underpinnings is further supported by a recent lesion study on subjects with focal lesion to left TP, right TP or outside TP [246].

Using a landmark recognition and naming task, the author found that landmark naming was significantly inferior in the left group. In a second experiment it was also found that participants with left TP lesions had impaired naming of famous faces, supporting the notion that left TP contain systems that are important for retrieving proper names for unique entities.

Lateralized processes in identification of specific exemplars compared to identification at the level of basic category have been investigated by Laeng et al. [130]. Using a picture-name verification task and presenting the stimuli tachistoscopically in one of the two later visual hemifields, Laeng et al. found that left hemisphere (LH) is specialized for classifying objects at the basic level, whereas the right hemisphere (RH) is specialized for classifying objects at the most specific level of abstraction, i.e. the level of unique identity.

The results of the Laeng's study are consistent with those of a study by Marsolek [148] in the context of visual form perception and identification which shows that the left hemisphere (LH) preferentially encodes general and abstract representations and prototypes, whereas the right hemisphere (RH) preferentially encodes exemplars.

Another study which directly compared the processing of familiar (at the exemplar level) and unfamiliar faces and building was conducted by Engst et al. [63] by recording event-related potentials in a priming repetition task. The study focused on two levels of the unique entity recognition process: the access to stored structural representations and to identity-specific semantic knowledge. The distinction between these two separate levels of processing is at the core of

the most highly accepted models of face recognition [35] and object recognition and have been related to different ERPs components.

In their study, Engst et al. focused on the early repetition effect (ERE/N250r) which has been proposed to indicate the access to stored structural knowledge and the late repetition effect (LRE/N400), a possible indicator of semantic knowledge access. The results showed that an ERE/N250r component was present for both familiar faces and familiar building. Moreover, the scalp topography were indistinguishable between faces and building. On the contrary, the late repetition effect (LRE/N400) displayed a very distinct category-specific scalp topography. This results is relevant because showed that semantic knowledge about persons and similarly unique non-face objects have separate representations in the brain.

In summary, converging evidence from neuropsychological and neuroimaging studies suggest that unique entities from different categories share common mechanisms of processing. In particular, these mechanisms seem involve the later stages of processing, i.e. the access to long-term memory representations that mediate entity identification and proper name retrieval. These results are in line with our hypothesis that entities from different categories involve differential pre-semantic processing prior to access a common system of stored knowledge of unique identity, i.e. the system of singular concepts.



## Chapter 5

# The Problem of Identity in Information Systems

The problem of identifying and tracing the identity of individuals is not exclusive of cognitive agents. An equivalent issue can be found in computer-based information systems used for managing and integrating information about entities.

According to our model of singular cognition, each act of the identification process - whether perceptual or conceptual - is mediated by a system of singular mental representations about unique individual entities. These cognitive structures, which function as tools for unique reference, allow us to manage different mental activities concerning individual entities. They provide the means for storing information about unique individuals, identifying and re-identifying them across time, integrating information about them from different sources and mediating the cognitive interoperability with other cognitive agents by ensuring the referential agreement between them in communication or other kinds of interaction.

In the same way, computer-based information systems which manage information about real-world entities need mechanisms to guarantee the effectiveness of equivalent processes (i.e. storage, retrieval, identification, integration and tracking) from a technological point of view. Consider, for example, the simple case of a database which store information about the employees of an organization. Each record of the database should store information about a unique individual and provide a mechanism to correctly and unambiguously access to that information. Ideally, redundancy in the records should be avoided, i.e. the database should not contain multiple references to the same entity. When a request about a specific individual is submitted, the system needs to identify

the corresponding record and return the correct information to the user (i.e. the information about the person intended by the user's query). The identification mechanism is also in play when the information within a record needs to be updated (i.e. by adding new information or by changing stored information) or a record must be deleted from the database.

From these considerations, it should be straightforward that the performance of these information systems and their interoperability with other systems strongly depends on their ability to uniquely identify (and re-identify) the entities represented within the system and outside the boundaries of it.

As the cognitive system uses mental representations that "stand for" actual objects in the real world, an information system - which can not have a physical access to objects of the real world - needs of some kind of mechanism which uniquely represent the real-world entities which are coded in the system.

The commonly practiced solution in this case is to provide a placeholder (i.e. an identifier) for each represented entity. Of course, since many different systems may represent information about the same real-world entities and identify them with different identifiers, a big effort is made recently to find solutions for achieving the information integration across multiple heterogeneous sources.

Not only information systems deal with very similar representation and identification problems to those of cognitive systems, but also the research on knowledge representation and integration in information systems was influenced by a bias toward high level categories similarly to what we observed for studies on human knowledge representation. Indeed, the largest part of the research effort has been made on the problem of 1) studying and designing high level conceptual representations (i.e. general categories and their organization in ontologies) or 2) designing methods for aligning and integrating heterogeneous representations (e.g. schema level integration, ontology alignment and integration). Only few studies focused on the ground of the ontological representations, that is on the entities which populate ontologies and the relations between them.

However, the idea that uniquely identifying entities is at the core of the problem of information integration across multiple heterogeneous systems is nowadays increasingly diffuse, representing also one of the main pillars of the Semantic Web.

In this section we focus on these new entity-centric approaches, reserving particular attention to the problem of identification in the Semantic Web.

## 5.1 Entity-level Information Integration

Describing the functioning dynamics of singular concepts, we noted that singular concepts can be merged when the cognitive agent comes to identify two entities

previously taken to be distinct. A very similar process in information systems is involved in entity-level (or data-level) information integration.

While schema-level integration deals with combining the general schemas (i.e. the format, structure, and organization of the data in a system) of two or more representation systems into a coherent and global view, entity-level information integration focuses on integrating information at the entity-level (i.e. the singular representations of individual entities). Entity level integration concerns with deciding whether two entity descriptions refer to the same individual (or entity) and with merging the two entity descriptions (deciding what to include in the joint entity description).

The problem has been largely investigated by the database community. In this context entity-level integration is the process of determining the correspondence between singular instances from more than one database.

This problem is known by the name of entity resolution, record linkage, object identification, de-duplication, merge/purge, data association, identity uncertainty, field matching problem, reference reconciliation, and others.

Entity-level integration is difficult in database integration because similar data entities in different databases may not have the same key (or identifier). For example, an employee may be uniquely identified by name in one database, and by social insurance number in another. Determining which employee instances in the two databases are the same is a complicated task if they do not share the same key. Entity identification has been defined as the process of determining the correspondence between object instances from more than one database [138]. Combining data instances involves entity identification (determining which instances are the same), and resolving attribute value conflicts (two different attribute values for the same attribute).

As many names have been assigned to the same problem, as several different approaches have been proposed in literature to address it. Here we give just an overview of the variety of approaches in this research field.

The entity resolution problem was first identified by Newcombe et al. [169], and given a statistical formulation by Fellegi and Sunter [65]. Most current approaches are variants of the Fellegi-Sunter model, in which entity resolution is viewed as a classification problem: given a vector of similarity scores between the attributes of two entities, classify it as “Match” or “Non-match” [16].

For example, Ganesh et al. [77] proposed the use of distances between attribute values as a measure of similarity between the records they represent. A generalization of the the Fellegi-Sunter model can be found in [224] which proposed a method based on the Markov logic.

Recently, a multiple classifier system approach has been proposed by [275]. The method applies several classification techniques drawn from statistical pat-

tern recognition, machine learning, and artificial neural networks to determine whether two records from different data sources represent the same real-world entity. The results of a first evaluation of the method showed a significant improvement compared to previous methods.

Other approaches treated the problem of entity resolution as a name-matching task and used the notion of string distance (i.e. a metric for measuring the amount of difference between two sequences of characters) to resolve the matching problem [164, 47].

An innovative method for data integration appeared in [215]. In this work, the authors attached context information to simple data values. For example, context information on a stock price including currency value and scaling factor. This context information was stored using defined names and values, so that comparisons between data values under different contexts was possible. A set of conversion functions were defined to convert from one context to another.

Other studies which addressed the problem of integrating information at the entity-level deal with *identity uncertainty*, *object identification* and *co-reference resolution* in natural language processing.

Identity uncertainty arises whenever entities in the data are not labeled with unique identifiers or when those identifiers may not be perceived perfectly. Identity uncertainty has been studied for example in *citation matching*, i.e. the problem of deciding which citations correspond to the same publication. Many approaches have been proposed to address this issue, many of which used machine learning techniques [179, 132].

Another problem concerning entity-level integration is dubbed object identification. The problem, in this case, is that data objects can exist in inconsistent text formats across several sources (e.g. the same restaurant can be referred differently in two different web sites). The first methods of object identification have required manual construction of object identification rules or mapping rules for determining the mappings between objects. More recent approaches used machine learning methods to derive automatically these rules in specific domains [244].

Finally, co-reference resolution is typically done with unstructured texts and deals with the problem to find the nouns, pronouns and phrases that refer to the same entity in a text. Co-reference resolution has often been performed by learning pairwise distance metrics between mentions [227] or using conditional probabilistic models [151].



## 5.2 Identity and Reference on the Semantic Web

The idea that the integration of information across heterogeneous resources can be addressed by means of mechanisms of identification of unique entities is one of the main pillars of the Semantic Web. The key feature of the Semantic Web is not its use of knowledge representation technologies like ontologies, but the introduction of these technologies to operate over Web resources<sup>1</sup> as defined by URIs. Information about resources is represented by means of a language, resource Description Framework (RDF), which is based upon the idea of making statements about resources in the form of subject-predicate-object expressions named RDF triples or RDF statements. Each resource in a RDF triple is identified by means of a URI.

The general idea is that if people use standard names for resources (URIs), then the integration of information from different distributed sources will happen smoothly and efficiently simply by using URI identity as a means for merging representations about the same entities. However, the solution proposed by the Semantic Web to extend the use of a URI (Uniform Resource Identifier) to identify not just web pages, but any resource on the Web [20, 19] have lead to many problems of reference, identity, and meaning [100], generating what has been dubbed the *Identity Crisis* of the Semantic Web [44].

According to the Semantic Web vision, in contrast to past practice that generally used URIs for web-pages, URIs could be given to resources traditionally thought of as not “on the Web” such as abstract concepts, people, monuments and so on. It seems, with this ever-expanding notion of a resource, that very different kinds of things are being described by a notion of a resource, including web-accessible resources, like a webpage, and resources that are not, like the Colosseum.

The guiding example is that instead of just visiting Tim Berners-Lee’s web page to retrieve a representation of Tim Berners-Lee via http, you could use the Semantic Web to make statements about Tim himself, such as where he was born or the color of his eyes.

However, this solution to talk about anything with URIs leads to identification problems.

The first problem concerns the following question: What does a URI identify? For web pages or documents it’s pretty easy to tell what a URI identifies. The URI identifies the information that one gets when one accesses the URI

---

<sup>1</sup>Tim Berners-Lee, who originally expressed the vision of the Semantic Web, defined a resource as “anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., “to day’s weather report for Los Angeles”), and a collection of other resources. Not all resources are network retrievable; e.g., human beings, corporations, and bound books in a library can also be considered resources”.

with whatever operations are allowed by the scheme of the URI. Therefore, unlike names in natural language, URIs often imply the potential possession of whatever representations the URI gives one access to. But when a URI is used to identify something that is “not on the Web”, what a URI identifies or means is a question of use. A URI has no identity in of itself, but only in the context of its use. If the meaning of a URI is its use, then this use can easily change between applications, and nothing about the meaning (use) of a URI should be assumed to be invariant across applications.

The second identification problem, or *co-reference problem*<sup>2</sup>, deals with the fact that there is no way in the Semantic Web Vision to force people to use the same URI for identifying the same entity across different entity. Moreover, a single URI may be used to identify more than one resource. This leads to a proliferation of URIs which hinders significantly the integration of Semantic Web knowledge on the data level. Therefore if a reliable method for supporting the reuse of URIs for entities across application is not ensured, the risk is to produce, as noted by Bouquet et al. [28], “an archipelago of semantic islands where conceptual knowledge may (or may not) be integrated (it depends on how we choose the names of classes and properties, and on the availability of crossontology mappings), but ground knowledge is completely disconnected”.

Different solutions have been proposed in the Semantic Web to address the identification problems described above, with the common aim to allow information integration across systems and applications. A notable approach is the effort of the Linking Open Data Initiative<sup>3</sup>, which has the goal to “connect related data that was not previously linked”. The main approach pursued by the initiative is to establish owl:sameAs statements (meaning “this resource is the same as that resource”) between resources in RDF in order to resolve the co-reference problem of the Semantic Web. More precisely, the semantics of owl:sameAs dictates that all the URIs linked with this predicate have the same identity, implying that the subject and object must be the same resource. The problem is that in many cases one can only be sure that two URIs are equivalent within the confines of a specific application, whereas owl:sameAs asserts that two references are always the same. Therefore the major disadvantage with this approach is that the two URIs become indistinguishable by means of the owl:sameAs link, even though they may refer to different entities according to the context in which they are used. Moreover, the owl:sameAs approach in fact not address the problem of multiple identifiers for the same entity, in turn it supports their proliferation.

---

<sup>2</sup>Co-reference deals with ensuring that two different entities do not share the same name or identifier, and conversely identifying when two identifiers refer to the same entity. In the context of the Semantic Web we are therefore concerned with URIs.

<sup>3</sup><http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

A centralized solution for the problem of proliferation of identifiers has been proposed by [85]. The authors implemented a Coreference Resolution Service (CRS) to facilitate rigorous management of URI co-reference data, and enable interoperation between multiple Linked Open Data sources. The system is based on the idea of maintaining sets of equivalent URIs. Equivalent URIs are conceptually stored in a “bundle” - a set of identifiers referring to resources which are considered to be the same in a given context. A URI can exist in at most one bundle within a CRS instance. One URI in each bundle is nominated to be a canonical identifier, or canon, for that bundle, representing a “preferred” URI for the set of duplicates. An application that wishes to use data from multiple sources as if they were a single resource can process results by looking up URIs in a CRS and replacing them with their canons on the fly, reducing the multiplicity of identifiers to a single definitive URI.

### **5.3 An Entity-Centric System for tracing the identity of entities on the Semantic Web**

A different solution to the entity identification problem has been recently proposed within the OKKAM project [28, 29]. We pay particular attention to this solution because it shows the strong parallelism between the identification needs in a cognitive system and those concerning an entity-centric system. Moreover, the analysis of the functioning dynamics of an entity-centric system reveals that many aspects which govern the functioning of singular concepts have a counterpart in the functioning of entity-centric systems. This parallelism leads us to investigate possible contributions which a cognitive study on the identification problem may provide to answer analogous questions in a technological context and inspire possible solutions to some of the most crucial issues about entity identification in entity-based systems.

The goal of OKKAM is to develop an Entity Name System (ENS) for the (Semantic) Web, a web-scale infrastructure which can make sure that the same entity is referred to through the same URI across any type of content, format, application.

The key idea behind the proposal of an ENS is that the Semantic Web can become an open and scalable space for publishing knowledge only if there will be a reliable support for the reuse of URIs. The ENS has a repository for storing entity identifiers along with some small amount of descriptive information for each entity. When a request for an entity is submitted, the ENS decides if a URI for this entity is already available in the repository; if it is, then the ENS will return its URI, otherwise it will issue a new URI which will be stored in the

repository. In this sense the ENS is different from the Coreference Resolution Service described above [85]. Instead of creating a RDF repository in which the same real-world entity is denoted by two or more different URIs, and then trying to reconcile these URIs, the aim of the ENS is enabling any application which produces RDF content to reuse a globally unique URI for that resource. Instead of using one of the many possible names for an entity, the ENS provides a unique name (i.e. global identifier) for that entity. This leads to the possibility to relate and integrate - without additional efforts - all the contents referring to the specific entity through its unique global identifier.

The development of an ENS leads inevitably to issues of entity representation and identification (some of these issues are discussed in [235, 29]) .

The first issue deals with the representation and identification of an entity in the ENS. Since the system has to decide which (if any) URI in the repository corresponds to a given request, the question is how an entity is supposed to be described in a way that sufficiently distinguishes it from all other entities.

A possible solution is to define the type of an entity with a possibly associated schema for its representation (description). An entity repository with a strong notion of typing is expected to increase efficiency and effectiveness of entity identifier retrieval, because entities can be managed in virtually or physically separate repositories according to their types, and type-specific matching approaches can be implemented. This raises the challenge of finding the right granularity and the right set of types for organizing the repositories. The solution adopted in the ENS is the use of high level entity types, such as “person”, “organization”, “event” to provide an upper level organization of the entity profiles in the system.

A second, although related topic, is the use of schemata for the representation of the entities. Since the ENS is not a repository of information about entities, the idea is not to collect as much information as possible about each entity, but simply to provide a schema which should include the attributes that are most adequate for the identification of the respective entity. The envisioned solution in this case is the use of a core schema of attributes dynamically adapted, based on data learned from the usage of the entity repository. This core schema should guide storage of new entities and matching for candidates retrieval. However, for the usability in different situations the user is allowed to use the attributes he has at hand for querying the entity repository. Therefore, the system is supposed to cope with the translation of incoming queries into core schema requests (schema-mapping).

The representation issues described above provides a first point of contact between cognitive and technological issues. In particular the question of which attributes should be included in a core schema to provide an effective descrip-

tion to uniquely identify entities in the system is not trivial. We suggest that the study of how people organize information within individual concepts could provide interesting insight to answer this question. In the same vein, studying how people search for information about individual entities (for example using a search engine like Google) is another way to explore the same issue from a different perspective. The searching task can be seen as a way to pick up few relevant attributes from the singular concept of an entity and use them to identify the entity in that particular context. A system which is designed to consider the real ways of identification of its users is expected to increase efficiency and effectiveness in retrieval, as well as to incorporate important features in terms of usability.

Another issue related to the development of an ENS concerns the repository maintenance. The same mechanisms which we have described about the functioning of singular concepts have a technological counterpart in the maintenance of the repository of an entity-centric system. Following the classification which we have prosed in 2.2 we can identify four main processes.

- *Initiation*: The creation of a new entity in the system starts with an entity request. Given a request of an entity, the system decide if a URI for this entity is already available in the entity repository (using some method(s) for entity matching); if it is, then the ENS will return its URI (or at least a ranked list of candidates), otherwise it will issue a new URI which will be stored in the ENS repository with the information that was provided as part of the respective request. This process leads to create a new representation, i.e. entity profile, in the ENS.
- *Updating*: The information about an entity can be managed by updating and extending the information contained in the initial entity profile of the entity. For the update and extension of the information managed for the individual entities, different processes are involved: a) new information that is provided when further requests for the same entity are encountered is added to the initial profile; b) the usefulness of the stored entity information is analyzed and eventually the information is filtered, c) the age of the information is considered and old information can be deleted d) (only in rare cases) manual change of entity information via adequate user interfaces can be allowed.
- *Merging*: As a consequence of the fact that new profiles of entities are continuously added to the system and entity representation change incrementally as a consequence of the updating process, the system is supposed to revisit its identity decisions, i.e. it has to check if given the current status of information in the repository, entity matching would still support

the same entity identity decisions. As a result of such a process it might be detected that two entity representations (with separate identifiers) actually refer to the same real world entity, requiring corrective actions which produce a unified representation from the two initially separated profiles. This process is named entity merging.

- *Splitting*: The opposite case happens when the revision process detects evidences for the fact that two real world entities have been by mistake or lack of sufficient information been marked as identical. In this case the initial profile must be split into two different profiles and the information contained in it must be correctly divided up in the two profiles, by means of a process named entity splitting.

The repository maintenance offers the most clear example of the correspondences between the managing of singular representation in memory and that of singular profiles in an entity-centric repository, such as the ENS. In particular, the maintenance process implies the notion of tracing the identity across time and change. As we have noted, this is a fundamental aspect in singular cognition. Again, we argue that a study of this human ability from a cognitive perspective can reveal useful aspects that can be implemented in a technological model for entity management. Moreover, given that the system interfaces with human users (as well as with machines), a better understanding of the identification processes in play when people interact with the system should contribute to suggest technological solutions based on the strategies and needs of its users.

## Part II

# Novel Contributions





## Chapter 6

# The Entry Point in the Identification of Individuals

Humans have an extraordinary ability to identify objects in a very efficient way. Any individual object can be identified at multiple levels of abstraction. So, for example, whereas a dog can be identified as a “dog” (basic level), the same dog can be identified more generally as “animal” (superordinate level) or more specifically as “poodle” or “Fido” (subordinate or unique level, respectively). The first phase of our research aims to investigate whether there is a preferential level of abstraction at which an individual is first identified. Do we first identify our dog as “Fido” or as a “dog”? Is there a direct access to the identity node of Fido (i.e. the singular concept of Fido) during the identification process, or is the access mediated by higher level conceptual representations (i.e. general concepts)? These questions deal with the bottom-up access to singular concepts of individuals, that is the way through which the perceptual stimulus makes contact with its singular concept.

The aim of this study was to investigate whether people identify individual artifacts from three different categories (i.e. artwork, building and product) more frequently and as quickly (or more quickly) at the unique level (e.g. Mona Lisa) as at the basic level (e.g. painting), and whether they have direct (i.e. unmediated by higher level representations) access to the visual representations of these individuals at the level of unique identity.

To this purpose, we conducted three experiments. In a first experiment, it was tested whether individual artifacts as opposed to non-individual artifacts were most frequently named with unique level category names, compared to basic or superordinate category names. The second experiment, investigated whether individual artifacts were recognized faster than non-individual artifacts

at the unique level of abstraction. Finally, in a identity-priming experiment we studied whether unique-level primes facilitated the matching responses of individual artifacts but not of non-individual artifacts.

## 6.1 Introduction

In their seminal studies, Rosch et al. investigated “the principles by which humans divide up the world” [198, p.382]. The authors found that, although all objects can be categorized at different levels of abstraction, there is one level, called *the basic level*, that has a special status in categorization (a phenomenon known as *basic level advantage*). This means that of all the various categories to which a given item belongs (e.g. “poodle” “dog”, “mammal”, “animal”, “pet”), some appear to be more readily accessible to the human mind than others. Rosch et al. have described four operational definitions that summarize their *structural view*<sup>1</sup> of the basic level. 1. Basic-level categories are the most inclusive categories for which clusters of co-occurring attributes are listed. 2. Members of a basic-level category share similar overall shape. 3. The basic level is the most inclusive level at which highly similar sequences of motor movements are used to interact with objects in the class; 4. the basic level is the most general level at which an averaged shape of an object may be correctly identified as that object. According to this structural view, the advantage for the basic level arises because the basic level is the level at which objects show the largest gain in structural similarity.

To test the relation between basic level advantage and object identification, Rosch and colleagues [160, 198] used several object-identification tasks. Among other things, the authors found that in an object naming task, where participants were asked to name an object with the first word that comes in mind, people prefer to use basic-level terms to identify objects (e.g. “dog”) over more general or specific terms (e.g. “animal” or “poodle”). In a category verification task, where people were asked to verify whether an object is a member of a category, it was found that they were faster to verify objects at an intermediate level of specificity (such as “dog”) than at more general (i.e. “animal”) and

---

<sup>1</sup>Rosch and colleagues [199] originally suggested that basic level categories are special because they capture significant regularities or patterns in the features associated with these categories. According to Rosch [197] “a working assumption of the research on basic objects is that in the perceived world, information-rich bundles of perceptual and functional attributes occur that form natural discontinuities, and that basic cuts in categorization are made at these discontinuities” (p.31). We refer to this type of explanation of the basic level advantage as a structural theory, since it implies that certain categories are “basic” because of their structural properties, namely, the statistical associations between features and categories. It is important to note that, although Rosch’s research emphasizes structure in the world, she did not view this structure as existing independently of the human perceiver. Rosch was careful to explain that it is the interaction between the human perceiver and the world that specifies the basic level.

more specific (i.e. “greyhound”) levels. Finally, in an identity matching task, where participants were asked to judge if two stimuli simultaneously presented were physically identical, the matching responses were faster when participants were primed with a basic-level name rather than a superordinate name. Critically, subordinate-level names provided no additional priming over basic-level names. From these evidences, it has been argued that the basic level represent the *entry point* in object recognition, that is the level at which the object is first recognized.

In one influential account of basic-level effects, Jolicoeur, Gluck, and Kosslyn [119] explained the superiority of the basic level with reference to spreading-activation models derived from the Collins and Quillian’s model of memory [50, 49].

In spreading-activation theories, concepts (i.e. mental representations of categories) are stored in memory within a hierarchical structure. Concepts are assumed to be represented as nodes in the hierarchy, which are interconnected by class-inclusion propositions called ISA links. For instance, the knowledge that a poodle is a kind of dog is represented by connecting the node for poodle to the node for dog with a ISA link; knowledge that dogs are animals is stored by linking the dog node to the animal node, and so on. Other facts are stored as predicates attached to the various nodes. For instance, to store the information that a dog “barks”, the predicate “barks” is attached to the dog node; and to store the information that all animals “need nutrients to survive”, the predicate “need nutrients to survive” is attached to the animal node. The fundamental retrieval mechanism is spreading of activation. To make inferences about the properties of a given concept such as poodle, the model first retrieves all of the predicates stored directly with the corresponding node; but activation then spreads upward along the ISA links so that the predicates attached to more inclusive concepts also get attributed to the probe concept. For poodle, activation first spreads to the dog node, supporting the inference that the poodle barks, and then up to the animal node, supporting the inference that the poodle needs nutrients to survive.

Jolicoeur et al. [119] proposed that certain nodes within a Quillian like processing hierarchy serve as “entry points” for probing the semantic network. An entry point corresponds to the level where the perceptual stimulus first makes contact with its underlying memorial representation. Visual stimuli are first classified into one of these entry-level categories by means of a perceptual processing mechanism so that any information stored directly with the corresponding entry-level node becomes available earliest in processing. Additional information about the stimulus becomes available later, as activation spreads upward from the entry point toward more inclusive concepts or downward to-

ward more specific concepts. Basic-level effects are observed for typical category members because the basic-level category nodes serve as the entry-point for such items. For instance, a visual stimulus such as beagle first activates the dog node, providing rapid access to the name “dog” and other typical dog properties (e.g. has four legs and can bark). Retrieval of properties that the beagle shares with all animals takes longer as it requires a search of the semantic network upward from the entry point. Retrieval of properties idiosyncratic to the beagle takes longer either because nodes below the entry point must be searched or because more specific classification relies on finer visual details [48].

Research on human object identification demonstrated that the entry point could be modulated by at least two factors: typicality of an exemplar for its corresponding basic level and domain-specific expertise.

Concerning the former, Jolicoeur et al. [119] suggested that atypical category members fail to show a basic level advantage because their entry-points are specific rather than basic. An atypical member is structurally dissimilar to the other members of the same basic level category and therefore it is more easily categorized at subordinate levels than at the basic. For example, the entry level for a picture of a penguin would be the node corresponding to penguin rather than the bird node that serves as the entry point for more typical birds.

About the second point, expertise in a particular field is likely to shift entry level of many objects towards the subordinate level. The influence of experience on the entry point has been firstly emphasized by Rosch et al. [198]. For example, they reported an anecdote about one of their participants who was an expert airplane mechanic and seemed to recognize airplanes at a more subordinate level of abstraction relative to the rest of the participants tested.

Johnson and Mervis [118] studied the interaction of knowledge and basic-level categorization in individuals with varying levels of knowledge about songbirds. Results from a series of experiments showed that experience increased accessibility to categorical knowledge at subordinate levels, causing these levels to function as basic (that means an increases in the speed and efficiency with which subbasic-level information was accessed from semantic memory). However they found no evidence that the original basic level actually changed as a function of knowledge. It was never the case that experts responded significantly less quickly for trials involving category names at the basic level than for category names at subordinate level. Thus, the efficiency advantage of the previous basic level is not lost as knowledge about subbasic categories increases.

Consistent with this view, Tanaka and Taylor [240] found that experts are as fast to categorize objects from their domain of expertise at the subordinate level of abstraction as they are to recognize the same object at the basic level. A dog expert, for example, is able to recognize a picture of a greyhound as a

“greyhound” as quickly as he recognizes it as a “dog”.

In the domain of face perception, Tanaka [241] suggested a shift in human perception towards subordinate classification of familiar faces. The general idea is that even though few people are experts in recognition of objects from a particular category, all adults can be considered expert in human face recognition [239]. Therefore, if face recognition follows the pattern of other kinds of expert object recognition, people should show a downward shift in recognition as a result of experience. However, although both face expertise and object expertise promote increased access to levels of representation subordinate to the basic level, the subordinate level corresponds to different levels of abstractions. In the case of object expertise, the identification typically occurs at the species (“mastiff”) and subspecies (“neapolitan mastiff”) level of abstraction. In the case of face expertise, the level of abstraction corresponds to the most extreme subordinate level, that is the level of unique identity where the category label is a proper name referring to a single individual in the world (e.g., Barack Obama). Therefore, the face expertise hypothesis predicts that the entry point of face recognition is at the level of unique identity. Thus, a face will more likely be identified as “Barack Obama” rather than as a “person”. Converging evidence from four experiments supports the hypothesis that the entry point of face recognition is different from the entry point of non-face objects. For example, whereas common object were likely to be identified with basic-level names, familiar faces were more likely to be identified with unique identity names (i.e., proper names). In a category verification task, faces were verified as quickly at the subordinate level of unique identity as at the basic level. Finally, results from an identity-priming task shown that subordinate level proper names labels produced greater priming effect than the basic-level labels.

Similar downward shifts in recognition were found by Gauthier and Tarr [80] after participants were trained in identification of artificial objects (“Greebles”) specifically constrained to be similar to faces along several dimensions (e.g., similar features organized in similar configurations).

Further evidence that the entry point of object recognition can be different from the basic level, comes from a recent study by Belke et al. [17]. In this study the authors provides empirical evidence that art is distinguished from other real world objects in human cognition, in that art allows for a special representation in memory and identification based on artists’ specific stylistic appearances. Converging evidence from three experiments suggests that identification of visual art is at the subordinate level of the producing artist (e.g., participants matched a familiar painting with its artist’s name as fast as they matched it with the artistic genre).

## 6.2 Objectives and Rationale of the Study

The studies described above provide evidence that “for many objects (and perhaps many situations) we use identification routines at levels other than the basic level” ([119], p. 272). Dog experts, for example, identify dogs at species and subspecies levels of abstraction as quickly as at the basic level.

A special case is represented by human faces. In his study on face recognition, Tanaka [241] provided evidence that the entry point of familiar faces is shifted to the most specific level of abstraction (i.e., people first recognize faces at the level of unique identity), supporting the assumption (previously untested) of many models of face recognition that the first recognition occurs at the level of individual faces. Compared to other objects classes, faces seem to represent a special class of objects because they require the most extreme level of specificity in recognition in which an individual face is the only object of the category.

But is the level of unique identity the preferential access point to memorial representation exclusive for faces?

This question can be linked to the broader debate whether faces are processed by cognitive systems specialized for (and specific to) this particular class of stimuli or by more general cognitive systems, which are used for all objects (see [153] for a review of this literature).

This issue is particularly interesting as human faces seem to occupy a special status among other visual objects. The extraordinary skills of humans in recognize familiar faces may indicate the existence of specialized processing mechanisms unrelated to those involved in visual object processing.

Neuropsychological lesion studies on patients with acquired impairments of face recognition (a deficit known as prosopagnosia) have been often interpreted as evidence of the existence of specific processing mechanisms that are separate from those applied to other objects. The impressive double dissociation found in different studies (prosopagnosia without object agnosia [188] and objects agnosia without prosopagnosia [204]) would seem to indicate the existence of some face-specific processing systems. However, reported dissociations have been also interpreted in a different way.

The alternative explanation assumes that faces are not necessarily processed by specialized processing systems [242]. Faces would require more precise perceptual discriminations to distinguish between them because of their high inter-stimulus structural similarity compared to other objects. Therefore, a partial deficit to a common perceptual processing system would reveal itself more strikingly for those stimuli that require a greater degree of differentiation, namely faces. However, many studies have demonstrated that inversion is more detrimental to recognition of faces than objects [270] (a phenomenon known as inver-

sion effect) and that upright faces are recognized more holistically than objects [238], which led to the suggestion that faces are recognized using specialized visual mechanisms.

The debate over the extent to which face-processing engages specific modules has recently extended into the functional neuroimaging literature. The site of primary interest is the Fusiform Face Area (FFA) that has been found to be preferentially activated by face stimuli compared to other object classes [123]. Since the location of this area is consistent with the lesion site in prosopagnosic patients and it reflects two classical markers of face processing (holistic processing [210] and inversion effect [274]), it has been suggested that FFA is a locus of face-specific processing. Contrary to this position, a series of studies have shown that the putative role of FFA as “face area” may be the result of our extensive experience with faces. For example, the two markers of face processing (holistic processing and inversion effect) have been obtained with non-face objects for expert subjects (e.g. dog experts). Brain imaging studies have shown that expertise recruits the FFA, increasing the response of this area to object of expertise compared with control objects.

From these evidences, it stands out that several questions about the mechanisms of face and object processing remain unresolved.

Up to now most of the research on this issue has focused on whether the cognitive and neural processes that are used for identifying faces are the same as or different from those that are used to recognize other kinds of objects. The common approach used in these studies contrast faces and objects in formally similar tasks and compare the effects of the same experimental manipulation. As an example of this approach, let’s consider a typical experiment of semantic priming. In this kind of experiment object naming and face naming are compared. In object naming, pictures of objects (e.g., an apple, a chair, a rabbit and so on) are preceded by semantically related pictures, whereas in face naming, familiar faces to be named are preceded by a face prime which is semantically related to the target. The facilitation effects in the two conditions are studied to make inferences about the organization of the underlying semantic representation of the two classes of stimuli. This example shows a common aspect of the current research on face and object recognition: the non-face stimuli used in these studies are accessed at the basic or (in some cases) at subordinate level of abstraction, whereas face stimuli are accessed at the unique level of identity. In other words, recognizing that an object is an apple involves the activation of a general concept (i.e., a cognitive representation of a category) whereas recognizing that an individual is Barack Obama involves the activation of a singular concept (i.e., a cognitive representation of a singular individual).

It remains unclear whether the differences between face and object processing

relate to different entry levels or to different recognition and categorization processes performed on the stimuli.

Therefore, the first aim of the present study is to investigate whether the entry point in the identification of unique non-face objects is at the level of unique identity as that of face objects [241].

Up to now research in the domain of object recognition has been concerned with object classes such as furniture, every-day-objects and even artificial objects, but very little is known about the representation and initial identification of unique entities belonging to these classes. For instance, what might be the first access to semantic memory when a person identifies the “Eiffel Tower”? If the entry point follows the Rosch et al.’s [198] structural definition, we should expect that the entry point in this case is at the level of “monument” or “tower” or even more general “work of art” corresponding to the basic level of the stimulus. According to the structural hypothesis, people may access to the unique level of identity only after the basic level is activated. Therefore, if the access to the subordinate level of identity is mediated through the basic-level, the structural account predicts that the basic-level categorization should be faster than the subordinate-level categorization. On the contrary, if the stimulus is recognized at the level of unique identity, as “Eiffel Tower”, recognition times should be as fast as or faster at this level than at the basic level.

Our hypothesis is that a person first recognizes an individual entity at the level of unique identity when she possesses an individual concept on that individual entity in semantic memory. We assume that the initial identification of an individual entity whose information is structured in memory as an individual concept yields cognitive processing that differs from that involved in the identification of objects which are not individuated in memory by means of individual concepts. Rephrasing the words of Rosch et al.[196], individual concepts follows the natural correlations and divisions of features distinguishing unique familiar entities and provide the entry points of recognition of these entities. Initializing the individual concept of an entity makes that entity unique and identifiable (i.e., atypical in a sense) from the other members of the same basic level category. Then this entity can be categorized faster at the most subordinate level of categorization, namely the unique level of identity.

We assume that having the singular concept of an object entails the direct recognition of the object through that concept. Therefore, the singular concept of an object acts as the access node to the knowledge that the agent has about the object. As a result, any information stored at the level of the singular concept becomes available earliest in processing.

We should note that our hypothesis finds support in several functional models of face recognition available in literature. These models suggest that the first



recognition of a known face occurs at the level of the individual face before to access more inclusive categorical knowledge.

In the Bruce and Young model [35], for example, the perception of a familiar face activates structural and view-independent long-term representations (face recognition units, FRUs). These FRUs are connected to amodal person identity nodes (PINs) which contain semantic-biographical information, such as occupation, hobbies, date of birth, etc.

The direct (i.e., non mediated by basic level knowledge) access to personal information stored in memory is also assumed by the interactive activation models of face recognition proposed by Burton, Bruce and Johnston [38] and by Bredart, Valentine, Calder and Gassi [36]. According to this models, PINs are activated from their corresponding FRUs, but they serve as modality-free gateway to stored personal information, coded at the semantic information units (SIUs).

We argue that the direct access to semantic information about unique individuals during the recognition process is not a cognitive process specialized for human faces, but is a general mechanism that humans use in the recognition process of unique identifiable entities.

Our strategy to test the hypothesis is to employ another category of unique entities, artifact, and predict that, if the entry point is set on the basis of the level of the uniqueness of the items within the category, the unique-level categorization of unique items should be faster than their upper-level categorizations.

### 6.3 Methodology

In this study, we explore whether the entry point in the identification of non-face entities can be shifted to subordinate levels of abstraction and whether this entry point is at the level of unique identity for those objects that can be recognized as unique and distinguishable from other objects (i.e, objects that have a singular concept in memory). To this purpose, we investigate in three experiments the initial identification of non-face objects belonging to three kinds of artifact types: artwork, building and product. Using tasks such as free naming, category verification and visual identity matching, paradigms were applied that had been predominantly used in the domain of object and face identification [198, 241, 17]. Performance on unique distinguishable entities from artwork and building classes (e.g. Mona Lisa, Eiffel Tower) is tested against unfamiliar objects from other artifactual categories which the entry point is expected to be at the basic level (as suggested by previous studies on object identification).

Entities from the third artifact class are used to investigate whether there is

a particular entry point in the identification of products that is different from other (artifact or non-artifact) objects.

Products represent a very special class of artifact from an ontological point of view. In a recent work on this subject, Vignolo [252] suggests that there are two ways to speak of products. One is the sense in which a product is referred as a model that can have many particular objects as its instances, the other is the sense in which a product is a specific instance of the product model. To clarify the distinction, consider the following example proposed in [252]. The car driven by Sean Connery in *Goldfinger* is an instance of the Aston Martin DB5. The Aston Martin DB5 of which the car driven by Sean Connery is an instance, is a model which might have many exemplars.

We hypothesize that this distinction may be reflected in semantic memory in two kinds of conceptual representations and presumably allows for a different type of processing. We note that in our hypothesis both types of representations are singular concepts. Claiming that the specific car driven by Sean Connery in *Goldfinger* has a corresponding singular memorial representation seem unproblematic. That particular car is processed at a conceptual level so specific that the car is in its class with no other members. However, the fact that the Aston Martin DB5 model corresponds to a singular concept in memory is less immediate. In this case many objects can fall under the same model class, nonetheless we argue that they are represented as a singular concept in memory and not as general concept.

To clarify this point, let's us consider the difference between a general category and a model category. We suggest that the difference lies in the graded structure of categories. Instead of being equivalent, the members of a category vary in how good an example (or in how typical) they are of their category [193, 200]. In the category of birds, a sparrow is generally considered a very typical exemplar of the category, a pigeon is moderately typical, whereas a penguin is atypical. Contrary to general categories, model categories do not present a graded structure. All objects falling under the category are equally good exemplars of the category. For instance, in the "iPod" model category, my iPod is not more typical than any other iPod belonging to the category. Then, many equivalent instances, which belong to the same model category, point to a unique model representation in memory. Therefore, the only difference between a singular concept and a model concept lies in the kind of relationship between instances and the conceptual representation. In the first case, there is a one to one relationship (a unique entity is represented by a singular concept), in the second case, there is a many to one relationship (many entities of the same model are represented by a singular model concept).

We propose that products are distinguishable from many other object classes,

since they can be first recognized at two subbasic levels of abstraction, that correspond to two kinds of singular concepts, one at the individual level (i.e., unique identity level), the other at the model level. In this sense, products provide a further test for investigating a downward shift in the entry point as a result of the level of specificity of the mental representation first linked to the item. In this study we focus on the model-level representation.

According to our hypothesis about the representation of identifiable artifacts (works of art, buildings and products), it is assumed that the entry point of these entities can be shifted to a subbasic level (i.e., the level of singular concept) and shows up at this level with highest frequency proportion in naming (experiment 1), with fastest category verification speed (experiment 2), with the largest amount of priming in visual identity matching task (experiment 3). Alternatively, according to the structural hypothesis proposed by Rosch et al. [198], artifacts should be identified first at a more general level (such as, for example, “building”, “artwork”, “product”, “bridge”, “portrait”, “audio player”, “car” and so on) and are more frequently named as such, are verified faster and yield higher priming gains on such basic level categories than at subordinate level categories.

For methodological reasons, the three experiments required participants being able to identify the critical stimuli (i.e., the stimuli used in the experimental conditions and contrasted with the control conditions) at the level of unique identity. For example, a person who has never encountered the statue of David by Michelangelo and who is not familiar with his name would neither be able to classify it as such in a naming task, nor to verify the David’s name in a category verification task, nor to respond to it in a priming task. Therefore, a procedure to omit from the analysis the items that can not be named at the unique level of identity is used in each experiment.

## 6.4 Experiment 1: Entity Naming

Every object belongs to more than a single category, but people must select only one when they are asked to name an object. In previous research, it has been shown that participants used basic-level names when asked to spontaneously name pictures of common objects [198, 119]. This result provides support to the notion that objects are first identified at the basic level of abstraction.

A free naming task was carried out in experiment 1 to test the hypothesis that people use subbasic level names (i.e., proper names or model names) when they are asked to freely name pictures of entities of which they have singular concepts in semantic memory.

Participants were shown with pictures of familiar entities (i.e., famous enti-

ties that can be commonly identified at the unique level of identity) from three different categories (artwork, building and product) and non familiar entities from three contrast categories (home furnishing, utensil and musical instrument). Unfamiliar stimuli contained sufficient detail to be identified at the subordinate level (e.g., “rocking chair”, “ upright piano”). The task was to name each object as fast as possible with the first noun that came instantaneously to mind.

Our hypothesis predicts that participants should use subordinate-level names (proper names or model names) when identifying pictures of familiar artifacts and basic-level names when identifying non familiar objects.

### 6.4.1 Method

#### Participants

18 people (11 male, 7 female) participated to the experiment. Mean age was 35.3 years (SD=2.45) ranging from 23 to 38 years. Each participant was tested individually and was not paid for participation.

#### Stimuli

The stimuli consisted of 48 pictures, half of which were from three artifactual category (artwork, building and product categories) and half from three contrasting categories (home furnishing, utensil and musical instrument). As famous artworks, some of the most well-known paintings and sculptures in art history were selected (e.g., Mona Lisa, Sunflowers, The Pietà). Famous buildings were selected from those used in [90] (e.g., Eiffel Tower, Twin Towers, Leaning Tower of Pisa). Finally, for the product category, we used some of the most popular models of vehicles and electronic devices in Italy (e.g., Fiat 500, Iphone). The complete lists of the familiar and unfamiliar stimuli used in experiment 1 are reported respectively in table 6.1 and table 6.2.

<b>Artwork</b>	<b>Buildings</b>	<b>Products</b>
Mona Lisa	Eiffel Tower	Fiat 500
The Last Supper	Leaning Tower of Pisa	Mini Cooper
Sunflowers	Golden Gate Bridge	Beetle
The Scream	Rialto Bridge	Fiat Panda
The Pietà	Twin Towers	iPod Nano
Discobolous	Empire State Building	Walkman
David	St. Peter’s Basilica	Black Barry
The Statue of Liberty	Milan Cathedral	iPhone

Table 6.1: List of familiar artifacts used in Experiment 1

Pictures were standardized with Adobe Photoshop to 336 (450 × 600 pixels) square centimeters with the original width-to-height ratio maintained.

Home furnishing	Utensil	Musical Instrument
rocking chair	wooden spoon	electric guitar
folding chair	tea spoon	acoustic guitar
desk lamp	bread knife	trombone
floor lamp	flick knife	clarinet
tea table	fry pan	bongo drum
dining table	saucepan	bass drum
four poster bed	nail scissors	grand piano
cot	garden scissors	upright piano

Table 6.2: List of unfamiliar artifacts used in Experiment 1

## Procedure

The experiment consisted of four practice and 48 experimental trials presented on a computer with a 15" monitor (resolution 1024×768). Participants were instructed that they would see a series of pictures of objects and their task was to name each of the stimuli as fast as possible with the first noun that comes to mind.

At the beginning of each trial a short instruction appeared on the screen reminding participants to “say the word that names the object as quickly as possible”. After a 2000 ms interval, the written instruction was replaced with a 800 ms blank screen, which was followed by a 2000 ms picture-stimulus (either a familiar or a non familiar item), which in turn was followed by another blank screen. After 1500 ms, the participants were asked to start next trial by pressing any key on the keyboard. The stimulus order was randomized with the restriction that pictures depicting famous objects from the same category are not presented on consecutive trials. The experimenter sat behind the participant and noted down the verbal responses for each experimental trial.

The experiment was implemented in Matlab using the Psychtoolbox-3. Viewing distance to the screen was approximately 70 cm.

To exclude the possibility that basic-level categories were used due to a lack of familiarity with subordinate level categories, at the end of the naming trials, participants were asked to identify each stimulus on a very specific (subordinate) level. For example, participants were asked to indicate the title of a painting or the model of a car. Pictures that could not be named at the subordinate level were omitted from the analysis for the corresponding participant. In this and the following experiments participants were tested individually.

### 6.4.2 Results

Before analyzing the data, all incorrect responses were eliminated according to the following two criteria. First, verbal classifications were excluded from analysis if a person could not name an object correctly at the subordinate level in the post-experimental task. In the case of familiar individual entities, this

task required labeling of the specific entity with its proper name or model name (unique level), or with any description showing that the entity was identified at subordinate level. For instance if the Last Supper could not be named as “The Last Supper” or as “a Leonardo’s painting”, the corresponding trial was omitted.

Second, if an object was named wrongly in the experiment, the response was considered as incorrect. For example, if the Golden Gate Bridge was labeled as “The Brooklyn Bridge”, the corresponding trial was eliminated. Given these exclusion criteria, participants responded to 94% of familiar unique objects and 97% of non familiar objects. Thus, participants were very familiar with the subordinate level terms of the objects. Finally, all correct responses were classified into four levels of abstraction (i.e., unique level, subordinate, basic and superordinate levels). For the familiar entities, naming responses such as “artwork”, “building” or “product” were classified as superordinate responses; “painting”, “bridge”, “car” etc. as basic level responses; descriptions such as “a Van Gogh’s painting” or “a famous bridge in San Francisco” as subordinate level responses, and finally proper names (i.e. titles, building names or model names) as unique level responses.

The dependent variable of interest was percentages of frequencies. Independent variables were categories (artwork, building, product, home furnishing, utensil and musical instrument), familiarity (i.e. familiar-object or non-familiar-object) and level of categorization (i.e., unique, subordinate, basic, or superordinate).

In table 6.3 we reported the percentages of frequencies for each category and for each level of abstraction. No verbal responses were given at the superordinate level for familiar entities and less than 1% of verbal classifications at superordinate level for unfamiliar entities. As expected, no verbal responses were obtained at the unique level for unfamiliar entities. Therefore, we collapsed the unique and subordinated-level responses into a general subordinate level of categorization. This procedure was used to compare familiar and unfamiliar entities at a level of abstraction that is subordinate to the basic level. Furthermore, given the lack of superordinate level responses, verbal responses were analyzed considering only two levels of abstraction (i.e. subordinate and basic levels) in subsequent analysis (aggregated data are shown in table 6.4).

To test for differences between the three categories of familiar entities, percentages of frequencies were submitted to two-way ANOVA with Category (artwork, building and product) and Level of Categorization (subordinate and basic) as within-participant factors. The main effect of object category was not significant,  $F(2,34)=2.09$ ,  $p=0.14$ , whereas the effect of level of categorization was significant,  $F(1,17)=67.64$ ,  $p<0.001$ . Category  $\times$  Level of Categorization

Category	Levels			
	Unique	Subordinate	Basic	Superordinate
Familiar Entities				
<i>Artwork</i>	0.85	0.03	0.12	0
<i>Building</i>	0.83	0.03	0.14	0
<i>Product</i>	0.66	0.15	0.21	0
Unfamiliar Entities				
<i>Home Furnishing</i>	0	0.47	0.52	0.01
<i>Utensil</i>	0	0.48	0.52	0
<i>Musical Instrument</i>	0	0.31	0.68	0.01

Table 6.3: Percentage Frequencies by Object category and Level of abstraction

Category	Levels	
	Subordinate	Basic
Familiar Entities		
<i>Artwork</i>	0.88	0.12
<i>Building</i>	0.86	0.14
<i>Product</i>	0.80	0.20
Unfamiliar Entities		
<i>Home Furnishing</i>	0.48	0.52
<i>Utensil</i>	0.47	0.53
<i>Musical Instrument</i>	0.32	0.68

Table 6.4: Percentage Frequencies by Object category and Level of abstraction: aggregated responses

interaction was also not significant  $F(2,34)=1.42$ ,  $p=0.25$ .

The same analysis was conducted to test differences among the three categories of unfamiliar entities (i.e. home furnishing, utensil and musical instrument). As expected, neither the main effect of object category  $F(2,34)=0.48$ ,  $p=0.48$ , nor the Category  $\times$  Level of Categorization interaction was significant,  $F(2,34)=3.06$ ,  $p=0.07$ . The effect of level of categorization was also not significant,  $F(1,17)=2.23$ ,  $p=0.15$ .

Given the lack of the main effect for category and interaction, the three categories of familiar entities were collapsed to obtain individual frequency scores for the domain of familiar objects. The same procedure was used for the three categories of unfamiliar entities, by collapsing the subordinate and basic-level responses across the categories.

Responses were collapsed and analyzed across participants. Figure 6.1 shows the percentages of subordinate and basic level responses for familiar and non-familiar objects. Participants named familiar entities in 90% of the trials at the subordinate-level and 16% at the basic-level. Unfamiliar entities were identified with subordinate-level terms on 43% of the trials and with basic-level terms on 58% of the trials.

A  $2 \times 2$  repeated-measures analysis of variance was performed with Fa-

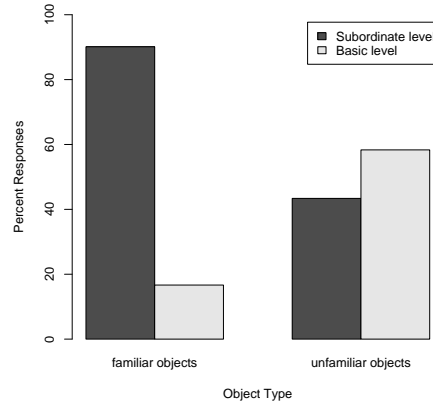


Figure 6.1: Percentage of basic level and subordinate level labels used in the naming task.

miliarity (familiar entities, non-familiar entities) and Level of Categorization (Subordinate and Basic levels). The analysis shows that the effect of Category was not significant  $F(1, 17) < 1$ ,  $p=0.93$ . On the contrary, the Level of Categorization was significant  $F(1, 17) = 12.03$ ,  $p < 0.01$ , as well as the Category  $\times$  Level of Categorization interaction,  $F(1,17)=76.11$ ,  $p < 0.001$ . The interaction (see Figure 6.2) indicated that familiar entities were more frequently named with subordinate-level terms than were unfamiliar entities, whereas unfamiliar entities were more frequently named at the basic level than familiar entities.

In a more detailed analysis, we also investigated the nature of concepts participants applied in naming of familiar entities. From table 6.3 we can note that familiar entities were predominantly identified at the unique level of abstraction. This kind of identification comprised using the title of an artwork, the name of a building or the model name of a product. However, we found that in some cases additional information were added by participants to identify the target entity. In particular, the artist's name were used in addition to the title of an artwork (e.g. Sunflower by Van Gogh), the name of the city in addition to the name of a building (e.g. St. Peter's Basilica Rome) and the brand in addition to the model name of a product (e.g. Fiat Panda). In the case of artwork, the represented object is also used in some descriptions. To perform the comparison between these two kinds of identification, we distinguished between *narrow unique level identification* (i.e. identification by proper names) and *broad unique level identification* (i.e. identification by proper names with additional information). This analysis was carried out independently of category. In total, participants used 88% of narrow identifications and 12% of broad identifications



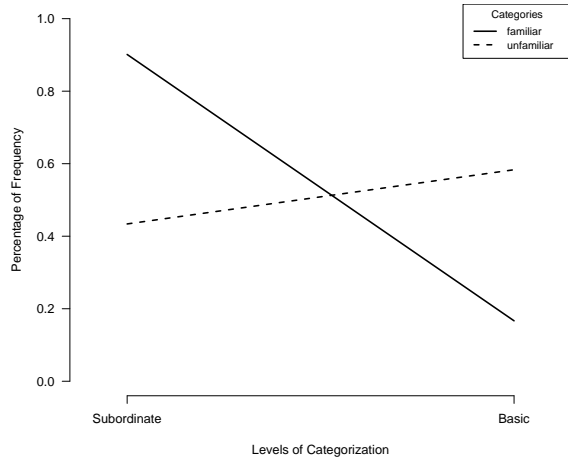


Figure 6.2: Category  $\times$  Level of Categorization interaction.

( $p < 0.001$ ). From this analysis it appears that using proper names were the predominant familiar entity identification.

Furthermore, restricting the analysis to the artwork category, we investigated the specific kind of art-related classification which were proposed by Belke et al. [17]. Among the art-related classifications, comprising the artist’s name, the artistic style and the title of the artwork, the author found that art objects were most frequently named with the artist’s name. In contrast with this result, we found that the title was most frequently mentioned (0.86%), followed by the artist’s name (0.14%). The artistic genre was never used by participants to identify artwork. In this respect, we didn’t confirm the assumption that art-objects allow for a special kind of identification based on individual artists’ styles that may serve as an entry point in recognition. On the contrary, we found that the participants’ naming behavior was not different for artworks compared to the other individual objects tested in the experiment.

To summarize, the results of the experiment 1 showed that entity naming of familiar objects differed from unfamiliar objects. Consistent with our hypothesis, unfamiliar objects were identified at a more general level of abstraction (basic level), while familiar objects were named at a more specific level (subordinate level). Results are in accordance with the assumption that familiar entities allow for a preferential level of identification which corresponds to the most specific level of unique identity. This result is analogous to the previous findings in the domain of face recognition [241] and art recognition [17] in which people preferred subordinate-level names over basic-level names to iden-

tify respectively faces and art objects. However, in contrast with the results by Belke [17], we found that art objects as other kinds of familiar entities are preferentially identified by the artwork’ title instead of by the artist’s name. As suggested by Tanaka [241] for face naming, it is possible that present results do not reflect an increased accessibility of the unique identity representation but instead naming preferences of people or social convention which encourage the use of proper names to refer to familiar entities. According to this interpretation, it is still possible that people identify familiar entities at the basic level, but choose to name them with the commonly used proper names. For this reason, the results of the experiment 1 show only a preference for this level of identification but are not sufficient to make inferences about the preferential access to semantic memory. In experiment 2, the accessibility to memory representations was directly tested analyzing reaction times in a category verification task.

## 6.5 Experiment 2: Category-Verification Task

In experiment 2, we used a category verification task similar to that adopted by Tanaka [241] in the domain of face recognition and by Belke [17] in art recognition. Participants were shown with a superordinate, basic or subordinate level category name and a brief time later were shown with a picture. Their task was to indicate whether the picture was an exemplar of that category. The results were compared between familiar and unfamiliar objects, selected from the categories used in experiment 1 (i.e., artwork, building and product) for familiar entities, (home furnishing, utensil and musical instrument) for unfamiliar objects. The choice of stimuli items and selection of verbal categories was oriented on the findings of experiment 1.

In the experiment, participants were asked to verify exemplars from these categories at superordinate (e.g., “artwork”, “building”, “furnishing”), basic (e.g., “painting”, “tower”, “chair”) and subordinate levels (e.g., “Mona Lisa”, “Eiffel Tower”, “rocking chair”) of categorizations.

In previous research [198, 119], it has been shown that participants were faster to categorize exemplars at the basic level (e.g., verifying that an entity is a “dog”) than categorizing exemplars at the superordinate level (e.g., verifying that an entity is an “animal”) and at the subordinate level (e.g., verifying that an entity is a “poodle”). Therefore, according to the basic-first hypothesis, artifacts should be categorized first at the basic level (regardless of the fact that they are familiar or unfamiliar). That is, basic level verifications should be faster than superordinate verifications and than subordinate verifications (unique identity name or model name verifications). For instance, people should be faster to verify that a picture is a “painting” than to verify that it is an “artwork” or

“Mona Lisa”.

On the contrary, we expect that subordinate-level representations will be more accessible than the basic-level representations for familiar objects. That is, participants should be as fast or faster to verify the unique identity of a familiar object (e.g., “Mona Lisa”) than to verify that the object is an “artwork” or a “painting”. We expect the same pattern of results for products, like familiar car models. That is, people should be as fast or faster to verify that a car is a “Fiat 500” than to verify that is a “vehicle” or a “car”.

### 6.5.1 Method

#### Participants

Twenty participants took part in the experiment. Mean age was 31.15 (SD=6.35), ranging from 23 to 45 years. Participants were tested individually and they were not paid for participation.

#### Stimuli

Pictures were chosen from the categories used in experiment 1 (artwork, building, product, home furnishing, utensil and musical instrument). The choice of stimuli items and the selection of verbal categories were oriented on the findings of experiment 1. For each category we selected 4 items. Additionally, four pictures other than those used for experimental trials were selected as practice trials.

#### Procedure

At the beginning of the experimental session, participants were presented with instructions explaining the category verification task on a monitor screen. They were also provided with the complete list of the subordinate-level terms for all of the 24 target exemplars presented in a random order one after the other. Subsequently, to signal the beginning of each trial, a fixation cross appeared for 1000 ms on the monitor. Next, a blank screen appeared for 1000 ms, followed by a category word which remains for 2500 ms. Finally, after 500 ms blank interval, the category name was replaced with a picture. The participants’ task was to verify whether the picture matched the category name, by pressing as quickly as possible the corresponding TRUE or FALSE buttons. The picture remained on the screen until the answer was given. The two response keys were counterbalanced for hand across participants. Trial order was fully randomized. Figure 6.3 illustrates the design of a sample trial in the category-verification task used in the experiment.

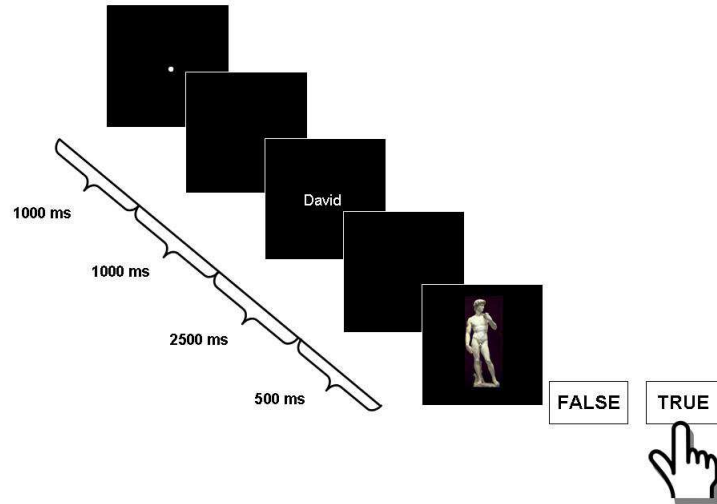


Figure 6.3: Trial presentation sequence in the category verification task. On each trial, a word was viewed (at superordinate, basic or subordinate level), followed by a picture, and the subjects were asked to indicate whether the picture matched the word.

The experiment consisted of 144 experimental trials, resulting from 24 items with two response types (TRUE and FALSE) and three levels of categorizations. That is, each item was shown six times. In the superordinate level and true condition, the category-word could be “artwork”, “building”, “product”, “furnishing”, “utensil”, “musical instruments”. In the basic level and true condition it could be “painting”, “tower”, “phone” and so on. Finally in the subordinate level and true condition the category word was the proper name of the artifact, the model name of the product or the specific type of furnishing, utensil or musical instrument. In the false conditions, category words were taken from a different exemplar of the same higher-order level category. For example, the “Eiffel Tower” letter string and the “Leaning Tower of Pisa” picture stimulus were paired, falling both under the same inclusive category “tower”. In the basic level condition, a false word label that shared the same superordinate category was provided (e.g., the letter string “painting” was presented with a “statue” picture stimulus, with both referring to the superordinate category “artwork”). False trials were designed with the restriction that each word-picture combination at the subordinate level would appear only once during the experiment and each word within a level of categorization would appear with the same frequency in order to prevent response bias. The experiment was implemented

in Matlab using the Psychtoolbox-3. The complete list of the category words used in the three categorization levels for true and false conditions is reported in Appendix A.1.

## 6.5.2 Results

Percentages of correct TRUE and correct FALSE responses by object category and level of categorization are reported in Table 6.5.

Category	Superordinate		Basic		Subordinate	
	CT (%)	CF (%)	CT (%)	CF (%)	CT (%)	CF (%)
Artwork	89	89	100	90	92	96
Building	85	86	92	93	93	92
Product	86	76	67	98	68	91
Home Furnishing	0.81	0.81	0.93	100	96	79
Utensil	0.81	0.81	0.93	100	0.81	0.81
Musical Instrument	100	100	0.90	100	0.84	0.67

Table 6.5: Percentage of correct TRUE (CT) and correct FALSE (CF) responses by category.

Aggregating the data by category, we found that participants, presented with familiar entities, correctly responded “true” on 83%, 86% and 84% of the trials for superordinate-level, basic-level and subordinate-level categorizations, respectively. For false trials, participants correctly responded on 98%, 94% and 93% of the trials for superordiante, basic and subordinate level categorizations, respectively. Responses to the unfamiliar entities, showed that participants correctly responded “true” to 88%, 92% and 94% of the trials for identifications at the superordinate level, basic level and subordinate level, respectively. For false trials, participants correctly responded “false” to 98%, 94% and 93% of the trials for identifications at the superordinate level, basic level and subordinate level, respectively.

An analysis of variance was performed on reaction times of correct true and separately of correct false responses. Before performing the analysis, trials with outlying RTs (i.e., below 300 ms or above 3000 ms) were excluded from the data set.

To test for differences between the three familiar categories mean RTs were submitted to two-way ANOVA with Category (artwork, building and product) and Category Level (superordinate, basic and subordinate) as within-participant factors. This analysis showed that the main effect of level of categorization was significant,  $F(2, 38) = 8.93$ ,  $p < 0.001$ . Neither the main effect of category  $F(2, 38) = 1.36$ ,  $p = 0.27$ , nor the interaction between category and category level were significant  $F(4, 76) = 0.20$ ,  $p = 0.93$  (see figure 6.4).

The same analysis was performed to test for differences among the unfamiliar categories. Mean RTs were subjected to a 3 (Category: home furnishing,

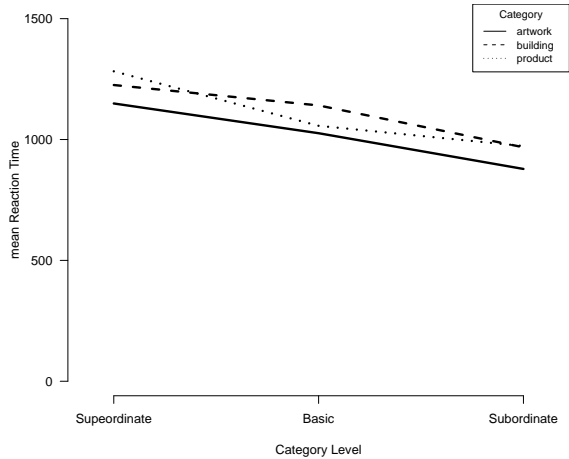


Figure 6.4: Mean Reaction Times for the three categories of familiar entities, at the superordinate, basic and subordinate levels in the true condition.

utensil and musical instrument)  $\times$  3 (Category Level: superordinate, basic and subordinate) within-participants ANOVA. As in the previous analysis we found that neither the main effect of category  $F(2, 38) = 1.03$ ,  $p = 0.36$ , nor the interaction between category and category level were significant  $F(4, 76) = 1.73$ ,  $p = 0.15$ . On the contrary, the main effect of level of categorization was significant,  $F(2, 38) = 11.20$ ,  $p < 0.001$ . The mean reaction times for the three categories of unfamiliar entities, at the superordinate, basic and subordinate levels are shown in figure 6.5.

Consequently, categories of familiar entities and categories of unfamiliar entities were collapsed to obtain individual mean RTs to familiar and unfamiliar entity types, respectively. Table 6.6 shows the separate reaction times for true responses as a function of category (Familiar vs. Unfamiliar) and category level (Superordinate, Basic and Subordinate).

Category	Category Level		
	Superordinate	Basic	Subordinate
Familiar	1200	1072	949
Unfamiliar	1236	979	1096

Table 6.6: Mean Reaction Times for the TRUE responses as a function of Category (familiar vs. unfamiliar) and Category Level (superordinate, basic and subordinate).

An analysis of variance (ANOVA) was performed for reaction times of correct true responses with Familiarity (familiar or unfamiliar) and Category Level

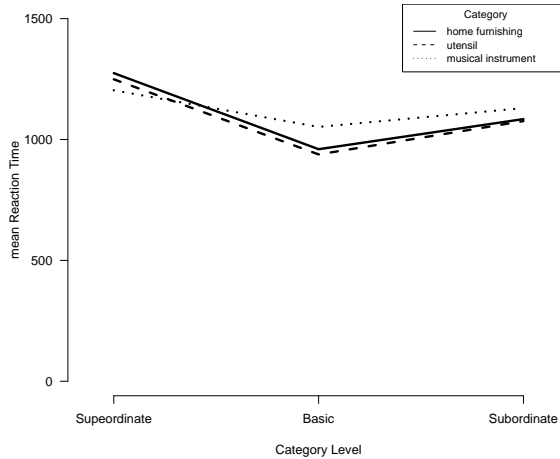


Figure 6.5: Mean Reaction Times for the three categories of unfamiliar entities, at the superordinate, basic and subordinate levels in the true condition.

(superordinate, basic and subordinate) as within participant factors. The main effect of Familiarity was not significant  $F(1, 19) = 0.93, p = 0.35$ , indicating that overall participants were not faster to categorize familiar entities than they were to categorize unfamiliar entities. On the contrary, the main effect of category level was significant,  $F(2, 38) = 13.61, p < 0.001$ . Critically, the Familiarity  $\times$  Category Level interaction was also significant,  $F(2, 38) = 5.69, p < 0.01$ . As shown in figure 6.6, participants were faster to categorize unfamiliar entities at the basic level than at subordinate level,  $F(1, 19) = 4.10, p < 0.05$ . For instance, they were faster to verify that a bread knife is a “knife” than they are to verify that it is a “bread knife”. On the contrary, for familiar entities, participants were faster to categorize entities at the subordinate level (i.e. unique level) than at the basic level,  $F(1, 19) = 7.72, p < 0.05$ . For example, participants were faster to verify that the David is “The David” than to verify that it is “a statue”. The results seem to confirm the assumption of a general basic-level advantage [198] for unfamiliar entities. However, contrary to this assumption, we found a different pattern of results for entities that can be identified at the unique level of identity (i.e. familiar entities). At the subordinate level (i.e. the unique level of identity) familiar entities were categorized faster than at the basic level, showing that the basic-level advantage disappears for entities that can be identified at the most specific level of identity.

Direct comparisons between TRUE judgments showed that subordinate-level judgments in the familiar category were significantly faster than subor-

dinate judgments in the unfamiliar category,  $t(19)=3.74$ ,  $p<0.01$ . The related comparison between reaction times for the familiar-basic and unfamiliar-basic categorizations showed the opposite pattern. Unfamiliar-basic judgments were significantly faster than familiar-basic judgments,  $t(19) = 2.36$ ,  $p<0.05$ .

In summary, these results demonstrated that familiar entities were identified differently from unfamiliar entities. People are faster to categorize familiar entities at subordinate level than they are to verify them at the basic level. On the contrary, verification times for unfamiliar entities were faster at the basic level than at the subordinate level.

The main contribution of this study is to support the assumption that the shift of the entry point in recognition towards the subordinate level is not peculiar of some special categories of entities but is a more general phenomenon concerning all the entities that have a unique representation in memory.

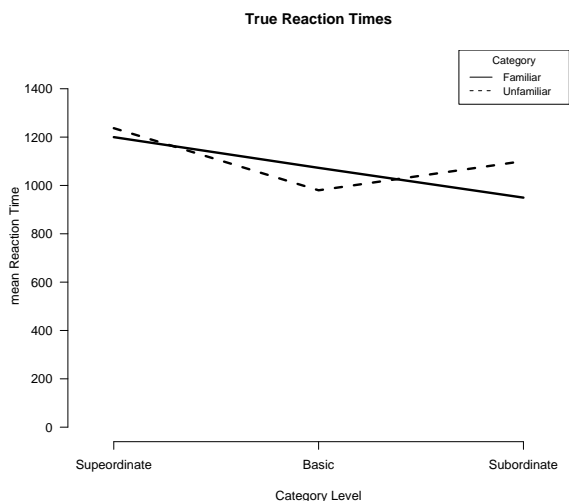


Figure 6.6: Mean Reaction Times for categorizing familiar and unfamiliar entities at superordinate, basic and subordinate levels in the TRUE condition.

An ANOVA was also performed for correct false reaction times with familiarity (familiar or unfamiliar) and category level (superordinate, basic and subordinate) as within-participant factors. Table 6.7 shows the separate reaction times for false responses as a function of category (Familiar vs. Unfamiliar) and category level (Superordinate, Basic and Subordinate).

The results of this analysis were globally in accordance with those obtained for correct true response times. In figure 6.7 we report mean reaction times for the correct falsification responses depending on familiarity and level of categorization. The main effect of familiarity was not significant,  $F(1, 19) = 1.40$ ,



Category	Category Level		
	Superordinate	Basic	Subordinate
Familiar	1108	1104	1052
Unfamiliar	1118	1010	1182

Table 6.7: Mean Reaction Times for the FALSE responses as a function of Category (familiar vs. unfamiliar) and Category Level (superordinate, basic and subordinate).

$p = 0.24$ . This means that people were not faster to verify familiar entities than unfamiliar entities. Instead, the main effect of level of categorization was significant,  $F(2, 38) = 12.97$ ,  $p < 0.001$ , indicating slower responses for a more specific level of categorization. Critically, the Familiarity  $\times$  Category Level interaction was also significant,  $F(2, 38) = 6.59$ ,  $p < 0.001$ . The interaction indicates that participants were faster to correctly reject unfamiliar entities at the basic level than at the subordinate level,  $F(1, 19) = 4.10$ ,  $p < 0.05$ , whereas they were equally faster to correctly reject familiar entities at basic level than at subordinate level,  $F(1, 19) = 0.161$ ,  $p = 0.69$ . The last result represents a difference compared to the previous analysis on the correct true reaction times. While participants were faster to verify a familiar entity at the subordinate level than at the basic level, they were equally fast to correctly reject a familiar entity at the subordinate-level as at the basic-level. This result could be explained arguing that the mismatch between the singular concept activated by the word category and that activated by the picture takes more time to be recognized. However, the result does not contrast our hypothesis since it shows that it is not the case that correctly rejecting a familiar entity at the basic-level is faster than rejecting a familiar entity an the subordinate level, as predicted by the basic-level advantage hypothesis. On the contrary the lack of a basic level advantage for the true rejecting trials of familiar entities indicated that representations of familiar entities are highly accessible at a specific level of abstraction which is related to the proper name of the entities.

As in the TRUE condition, we found that direct comparisons between FALSE judgments showed that basic-level judgments in the unfamiliar category were significantly faster than basic-level judgments in the familiar category,  $t(19)=4.07$ ,  $p<0.001$ . These results open the question whether a mechanism of inhibition may come into play to favor the access to singular representations compared to higher level representations. To answer this question future experiments should compare familiar and unfamiliar entities from the same categories to reduce as much as possible processing differences due to the category.

In conclusion, the results of the experiment 2 provided evidence in favor of our hypothesis that people are faster (or at least equally fast) to verify entities at the unique level than at higher levels of abstractions.

In terms of mental representations, this suggests that people have direct access to singular concepts and that this access is un-mediated by higher level conceptual representations, when they identify entities which are represented in memory by means of singular concepts.

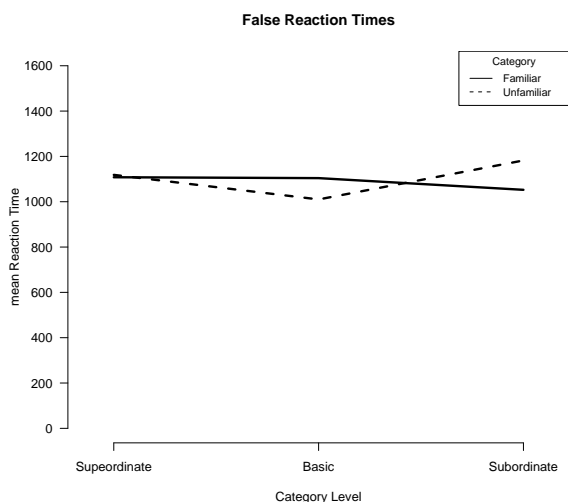


Figure 6.7: Mean Reaction Times for categorizing familiar and unfamiliar entities at superordinate, basic and subordinate levels in the FALSE condition.

## 6.6 Experiment 3: Identity Matching Task

Previous studies on the entry point issue, have shown that categorizations subordinate to the basic level require additional perceptual processing [119], as the identification at this level is based on more detailed perceptual discriminations during the initial processing. In experiment 3 such perceptual representations of individual artifact are directly examined using an identity matching task. Our prediction is that having a singular concept of an entity can favor the rapid perceptual analysis of information from that entity, because perceptual information may be part of the singular concept of the object. Participants are presented with a word prime (basic-level prime or subordinate-level prime) or a neutral prime (consisting of the letter string “blank”) followed by two simultaneously shown pictures. The participant’s task is to decide whether the two pictures are visually identical or different. Facilitation is measured by the difference in reaction times between primed and neutral trials for the same matching picture stimuli. The identity-priming paradigm assumes that the word prime activates the participant’s visual representation, which in turn is used to en-

hance the perceptual matching response [184, 198]. The stronger the priming is, the shorter the reaction times will be. The degree of facilitation depends on the match between mental representations, as elicited by the word stimulus and its correspondence with the physical picture stimulus.

This task has been first used by Rosch et al. [198] to examine the category level at which people represents objects. They found that relative to the neutral condition, basic-level words (e.g. “dog”) produced more facilitation than superordinate level words (e.g. “animal”). More important for the aim of the present study, it was shown that subordinate-level words (e.g. “poodle”) did not produce more facilitation, even though they convey more information about the visual appearance of objects. From this evidence, the authors argued that people represent most objects at the basic level of detail.

Instead, we hypothesize that a proper name or a model name should activate a unique identity representation in memory that would facilitate the visual comparison task for pictures representing the referent of these names. On the contrary, if participants represents unique artifacts, like general objects, at the basic level, no differences in facilitation should be found between subordinate-level primes and basic-level primes.

## 6.6.1 Method

### Participants

Fourteen participants (8 female) took part in the experiment. Mean age was 28.5 years (SD=5.17) ranging from 23 to 38 years. None of the participants participated in Experiments 1 or 2. Each participant was tested individually in a quiet room.

### Stimuli

Target stimuli consisted of four pictures selected from each of the four categories (artwork, building, product, home furnishing, utensil and musical instrument) used in the previous experiments, resulting in 24 picture stimuli. For different responses, each target picture was paired with a different picture which shared the same basic level of the target (e.g. two different paintings). The complete list of the paired stimuli used in the experiment is reported in Appendix A.2

### Procedure

At the beginning of the experimental session, participants were shown written instructions that explained the procedure for the identity matching task.

The participants' task consisted of judging (as fast as accurately as possible) whether two simultaneously presented stimuli were physically identical or different. Before starting the experiment, participants performed eight practice trials followed by 144 experimental trials. At the beginning of each trial, a ready signal consisting of a fixation cross appeared for 1000 ms in the center of the screen. After that the cross was replaced by a word prime or the neutral word "blank" for 2500 ms. Word primes were either basic-level words (i.e., "artwork", "building", "product" for familiar entities and "furnishing", "utensil", "musical instrument" for unfamiliar entities) or subordinate-level words (e.g., "Mona Lisa", "Eiffel Tower", "Iphone" and so on). Subsequently, a 300 ms blank screen interval was shown and then followed by the simultaneous appearance of two pictures. The two pictures remained on screen until participants pressed the key marked SAME (indicating that the two pictures were physically identical) or the key marked DIFFERENT (indicating that the two pictures were physically different). The trial presentation sequence is presented in Figure 6.8.

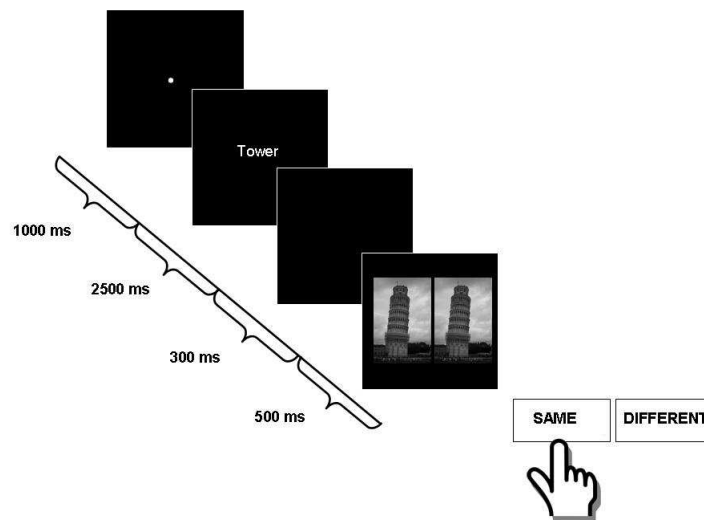


Figure 6.8: Trial presentation sequence in the identity matching task. On each trial, a word was viewed, followed by the simultaneous appearance of two pictures, and the subjects were asked to indicate whether the pictures were physically identical.

The two pictures presented in the "same" conditions were either two familiar objects from artwork, building or product categories, or two unfamiliar objects from furnishing, utensil or musical instrument categories. In the "different" conditions, the two pictures shared the same basic level (e.g., two different

paintings) with the restrictions that each combination appeared only once and all stimuli appeared with equal frequency. Thus, the three types of word primes, 24 target items, and two response types yielded a total of 144 experimental trials (see the Appendix A.2 for the complete list of experimental stimuli used in the experiment). Participants were instructed to answer by pressing one of two buttons corresponding to “same” and “different” answers. The two response keys were counterbalanced for hand across participants. Trial order was fully randomized. The experiment was implemented in Matlab using the Psychtoolbox-3 and run on a Dell Latitude D630 with a 15” monitor.

## 6.6.2 Results

Before performing the analysis, reaction times were adjusted by setting boundaries to eliminate outliers. The lower boundary was set to 200 ms and the upper boundary was set to 2000 ms, which corresponds to approximately 2.5 standard deviations from the mean ( $M_{RT} = 881$  ms,  $SD = 392$ ).

Table 6.8 shows the mean reaction times depending on prime level and object category.

Category	Priming		
	Neutral	Basic level	Subordinate level
Artwork	966 (44)	916 (52)	847 (49)
Building	970 (38)	936 (42)	860 (59)
Product	917 (49)	913 (54)	844 (50)
Home Furnishing	982 (76)	842 (40)	915 (58)
Utensil	996 (49)	866 (48)	927 (53)
Musical Instrument	946 (49)	899 (60)	932 (54)

Table 6.8: Mean RTs in milliseconds (and standard errors of the mean) by Object Category and Prime Type

In the following analysis, we used “same responses” to measure priming effects. To obtain priming scores, differences in reaction times were calculated between responses for correct same responses when pictures were preceded by a neutral word as compared with when they were preceded by either a basic-level word or a subordinate-level word. Mean priming scores for familiar and non-familiar objects were analyzed and compared.

To this purpose, we calculated a mean basic-level and subordinate priming score by averaging priming scores for the three categories of familiar objects and for the three categories of unfamiliar objects at the two levels of abstraction: basic and subordinate. Then, for each participant a mean priming score was calculated for each of the six target categories at the two levels of abstraction.

To test for differences between the three familiar classes of objects, we sub-

mitted priming scores to a two-way ANOVA with category (artwork, building and product) and priming condition (basic and subordinate level) as within-subjects factors. Neither the main effects of category,  $F(2,26)=0.19$ ,  $p=0.82$ , nor their interaction,  $F(2,26)=0.39$ ,  $p=0.67$ , was significant. Given of lack of difference between familiar categories, we collapsed the three categories of familiar entities (artwork, building and product) to obtain individual mean priming scores for a general category of familiar entities.

The same analysis was performed on the three contrast categories (home furnishing, utensil and musical instrument). As in the previous analysis, neither the effects of category,  $F(2,26)=1.81$ ,  $p=0.18$ , nor the interaction between category and level of abstraction was significant,  $F(2,26)=0.001$ ,  $p=0.99$ . Consequently, non-familiar object categories were collapsed to obtain individual mean priming scores for a general category of unfamiliar entities. This was done by averaging priming scores for basic and subordinate level responses across the three categories.

In Figure 6.9 we show the amount of total facilitation produced by the basic-level and subordinate-level primes depending on the object category (familiar vs. unfamiliar).

To test for differences in priming effects between familiar and unfamiliar objects, an ANOVA was performed with object domain (familiar or unfamiliar) and category level (basic or subordinate) as within participant factors. The analysis revealed a significant domain  $\times$  level of categorization interaction,  $F(1, 13) = 7.05$ ,  $p < 0.05$ . No other effects were significant. The interaction showed that additional priming effects were found at the subordinate level for familiar entities, but not for unfamiliar entities. The opposite pattern was found at the basic level, where additional priming was revealed for unfamiliar entities but not for familiar entities. However, a direct comparison between basic and subordinate-level primes revealed that the difference was significant only for familiar entities,  $p < 0.05$ . For unfamiliar entities the difference did not reach the 0.05 level of significance ( $p = 0.33$ ).

Thus, people recognized familiar entities faster at the unique level of identity and were able to access elaborated visual representations when primed with a matching entity proper name. According to the logic of the identity-matching paradigm, these findings suggest that for familiar entities, participants are able to activate unique-level visual representations, that are used to bear the identity matching task. We argue that such perceptual representations are part of the singular concepts of individual entities. Therefore the results suggest people have quick access to singular concepts during initial visual processing of individuals.

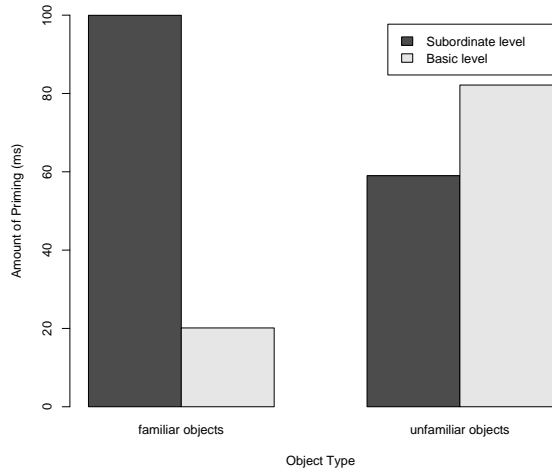


Figure 6.9: The results of Experiment 3 showing the amount of facilitation for basic and subordinate level primes for familiar and non-familiar objects.

## 6.7 General Discussion

The purpose of the study was to provide empirical evidence for the direct access to semantic memory of unique entities through individual concepts. Converging empirical evidence from three experiments, which have previously proved sensitive to address the object identification issue, suggested that the initial point of contact between the perceptual stimulus of a unique distinguishable object and its memory representation is not mediated by high level conceptual structures (i.e. general concepts).

Our study suggests that having an individual representation of an object in memory (i.e. individual concept) shifts the entry point of recognition to the most subordinate level of categorization, that is the unique level of identification. The recognition mechanism of unique familiar entities is different from that of entities that can not be identified at the unique level of identity (i.e. unfamiliar entities). In principle, a familiar individual could be first recognized as whatever other unfamiliar individual, namely as a member of a category (more likely as a member of a basic level category). Our experiments shown that this is not the case.

In a naming task (experiment 1), whereas common entities were likely to be identified with basic level category labels, it was found that familiar unique entities were more likely to be named with unique identity names. In a category verification task (experiment 2), we found that unique familiar entities were ver-

ified more quickly (or rejected as quickly as) at the subordinate level of unique identity than they were at the basic level. Finally, in experiment 3 the preferential access to a fine-grained visual representation for unique individual entities was demonstrated in an identity matching task. In this experiment, subordinate level primes (i.e. proper names) produced stronger RT facilitation compared to basic-level primes in matching pictures of familiar individual entities, but not in matching pairs of pictures of unfamiliar entities.

Considered together, these results suggest that whereas the entry point in recognition for most unfamiliar objects is at the basic level of categorization (i.e. the first contact with a memorial representation is at the level of a general concept), the entry point of unique familiar entities is at the subordinate level of unique identity (i.e. the first contact with a memorial representation is at the level of a singular concept).

The results of our study mirror previous findings in recognition of familiar faces [241] and visual art identification [17], in that a preferential accessibility to more specific representations in memory has been previously demonstrated for famous face and art recognition. However, in these studies the underlying idea is that there is something “special” in the target entities that lead people to develop specialized mechanisms of identification. Belke [17], for instance, explicitly argue that “art has a special status amongst external-world objects since it allows for a memorial representation based on stylistic features that are linked in semantic memory to the creating artist” (p.199). The special status of faces was instead conceived by Tanaka [241] in terms of expertise. Faces are different from other objects of expertise in that object expertise is a specialized activity that is achieved by relatively few individuals and only through explicit training. On the contrary, face expertise is a general ability that virtually all people possess (excluding people affected by rare disorders like prosopagnosia) and is acquired without training. According to the face expertise hypothesis, the high level of specialization in face recognition explains the shift of the entry point for faces at the most subordinate level of abstraction.

Expertise is also an important aspect of the study of Belke [17], in the sense that participants were people with a fairly good level of expertise in visual art. We believe that this methodological choice limits the generalizability of the results to people with limited art experience. The study leaves open the possibility that the organization of mental representations of visual art at the level of producing artist is a special feature of art experts. Even though we agree with the authors that art can be generally regarded as an expertise domain, because exposure to art, unlike encounters with every-day objects, is a rather limited event, we believe that it is still possible to investigate the identification mechanisms of art objects in people without relevant art-specific categories



acquired by training. To this purpose, in our study we used very famous art stimuli which are assumed to be part of common knowledge. The results of our experiments indicate that the specific level of unique identity is the level at which the recognition of famous art entities first occurs also in non experts. Contrary to the Belke's results, our experiments do not support the existence of a specialized mechanism for art identification at the level of the producing artist. In experiment 1, for example, the artist's name is used only in about 15% of the trials about artworks and always along with the title of the artwork (e.g. The Last Supper of Leonardo Da Vinci).

This evidence indicates that at least for very familiar art objects the entry point of identification is not dissimilar from that used for other classes of familiar entities. Consequently, the formation of style-based memorial representations could be a specialization which is acquired with the experience to deal with art entities whose memorial representations are not shaped by recurrent encounters and for which a singular concept has not yet been initialized.

We argue that the shift of the entry point toward the unique level of identity is not a peculiarity of a "special" category of objects, but is the general mechanism through which an entity recognizable at the level of unique identity is first recognized. In favor of this hypothesis, we did not find a significant difference between the three categories of familiar entities (artwork, building and product) used in our experiments.

This result is particularly interesting for a second reason. Converging evidence from the three experiments here described indicates that the identification of well identifiable products (e.g. the Beetle or the iPhone) is different from that of other generic objects (e.g. an unknown car or phone). The identification process of known products mirrors that of other familiar entities, in the sense that the first recognition occurs at the subordinate level of abstraction represented by the model name of the product. This means that, for instance, the iPhone is first recognized as "iPhone" rather than as a "phone". We argue that in this case the recognition is mediated by a memorial representation which has the same characteristics of a singular concept with the only difference that in this case many equivalent individual entities can be recognized by the same individual concept and referred with the same proper name. We remark that the model-based conceptual representation can not be considered as a general concept in that all the members of the category are equivalent: they are equally good exemplars of the category, they share the same relevant features and ultimately they have the same core meaning of the concept. Moreover, our study provides evidence that the model name is processed differently than the brand name in activating the corresponding conceptual representation. Previous studies [160, 119, 241] that used brand names to test the entry point of identification for products like

cars did not find a shift to the subordinate level of abstraction represented by brand names. On the contrary, we found that model names promote this shift. It is interesting to note that in their study on the neuropsychological status of brand names Gontijo et al. [86] found that brand names form a special lexical category that seem to occupy a somewhat intermediary lexical status between common nouns, nonwords and proper names. It would be interesting to investigate the neuropsychological and lexical status of model names to test whether they behave more similar to proper names than common nouns or brand names. The similarity between model names and proper names would be an evidence of their similar function in activating unique representations in memory.

In conclusion, our study provides some valuable insights to the current debate about the sequence of processing steps involved in visual object recognition. Traditional models of object recognition posit an intermediate stage between low-level visual processing and high-level object recognition at which the object is first segmented from the rest of the image before it is recognized [121, 58, 168]. However, other evidence suggests that object recognition may influence, and perhaps even precede, segmentation [182, 93]. Thus, the hypothesis which suggests that segmentation occurs prior to recognition, is currently subject to vigorous debate. Other evidence suggest that objects are perceptually categorized (e.g. bird) before they are identified at a finer grain (e.g. sparrow). Consistent with this second hypothesis, some behavioral evidence suggests that familiar objects are named faster at the basic level than the superordinate or subordinate level [198]. However, this is apparently not true for visually atypical members of a category [119]. Further, it has been suggested that visual expertise may lead experts to recognize stimuli from their expert category as fast at the subordinate level as the basic level [241]. Thus, the generality of the second hypothesis is also subject to debate.

The present study provides evidence in the same direction. The findings of three experiments challenge the hypothesis of a basic-level entry point of recognition of unique entities, where perceptual categorization precedes unique level identification and provide support for a direct (i.e. unmediated by general categories) access to unique information stored in individual concepts. Moreover, the results provide evidence in favor of the hypothesis that the shift of the entry point to the level of unique identity is not peculiar of “special” kinds of objects (like faces), but it is the general mechanism to access to individual-specific memorial representations.

## Chapter 7

# Associative and Semantic Priming in Recognition of Individuals

In chapter 6 we described three experiments which investigated the access point to singular concepts about individuals and we reported findings supporting the hypothesis of a direct - non mediated by higher level representations - bottom-up access to these concepts. In this chapter we explore another aspect of how our semantic representations of individuals are accessed and organized in memory. A priming experiment was conducted to investigate the relations between singular concepts, contrasting associative and semantic priming effects in a entity recognition task.

### 7.1 Introduction

Fast and reliable access to entity-specific information is of central importance in everyday life. Humans need to correctly store and retrieve knowledge about unique entities such as people, places, objects and other individual things relevant to their own existence. Our ability to recognize, identify and name all the entities which populate our environment and our life depends on this fundamental aspect of human cognition.

Although efforts to access this kind of information are subject to occasional incidents (see, for example, [272] for a description of errors in recognizing people), humans appear to be remarkably capable at storing and retrieving entity-related knowledge. For instance, when shown a picture of Barack Obama, people will know that he is a familiar person; they will be able to access biographical in-

formation, such as he is the 44th President of the United States, is married with Michelle Obama, represents the Democratic Party; and they probably will also be able to access to his name.

The way in which this information is structured in semantic memory and accessed from the perceptual input is a matter of considerable debate in cognitive research.

Many studies in literature focused on the problem to understand the organization of (and the access to) semantic memory for familiar people and several theoretical explanations have been proposed to explain person recognition and naming [35, 38, 36, 234].

However, little attempt has been made to compare the processes involved in person recognition and naming with those used for other kinds of entities. The only few studies (see for example [9, 55]) that have addressed this issue compared face and object processing at two different levels of abstraction. For instance, a common task used in these studies is the *face naming task*. In this task, participants are asked to name a known person by producing her proper name. To perform such a task, it is necessary to access to the specific memory representation of that unique person and retrieve part of it (in this case the proper name). On the contrary, in the corresponding task used for objects, namely the *object naming task*, a generic exemplar of an object category (e.g., a bottle) is shown and people are asked to name the object with the name of the category (that is a common name). When a person recognize an exemplar of the category “bottle” as “a bottle”, she does not access to the individual information of that specific bottle, but she assigns that exemplar to the bottle category and then she retrieves the name of the category. The findings of the experiments described in chapter 6 provided evidence in favor of a different access mechanism in these two cases.

As yet, very few studies [63, 90] have compared the organization of person-specific knowledge with that of other entities at the unique level of identity.

The primary rationale of the present research is to provide a contribution in this direction, comparing person and non-person entities that can likewise be accessed at the exemplar level in terms of recognition. More precisely, our aim is to investigate how our semantic representations of individual things are organized and accessed and how these representations are inter-linked with those of other individual things. To this purpose, we probe the semantic system using a priming experiment.

## 7.2 Categorical and associative relatedness between entities and priming effects

The analysis of priming effects has provided a feasible way of investigating the structure of semantic memory.

In this section, first, we will clarify the distinction between two different ways in which individual entities can be related in memory, namely by means of semantic and associative relationships. Then, we will explain how these relationships lead to different implications concerning priming effects (semantic and associative priming). Finally, we will review the literature on entity recognition and naming in the light of this distinction. Since our interest is on individual entities, we will reserve particular attention to the literature that addressed the problem of recognition and naming at the unique level of identity and in particular to the literature on face recognition and naming.

How semantic knowledge for individual entities is stored in long term memory is an open issue. One possible view - *categorical view* - is that semantic knowledge of individual entities has a categorical structure, as has been demonstrated to exist for generic objects [9, 110]. The idea is that memory representations of unique entities are interconnected by belonging to common categories. This view holds that the category “politician”, for example, exists as a node in a network and that all the exemplars of the category (e.g., Barack Obama, Bill Clinton, Nicolas Sarkozy) are connected to the corresponding node. The connection with the superordinate category creates an indirect link between these entities. An important implication of this view is that entities would be assumed to inherit the properties of the category to which they belong.

An alternative view - *associative view* - holds that the semantic knowledge for unique identifiable entities is different and that this knowledge is not structured according to categories. In this view, relationship between entities can be represented by networks of associative links but not by membership of a common category. According to this view, Barack Obama and Michelle Obama would be linked in memory because they are inter-connected by a directly associative factor (i.e. a partnership relationship). Moreover, it is assumed that knowledge of entities which are identifiable at the level of unique identity is individual and attributes cannot be automatically inferred from category membership. In the course of this section, we will discuss in more details the nature of categorical and associative connections between entities and we finally provide a clear definition of our use of these terms.

The two views described above imply different predictions about the priming effects that can be observed in experiments of entity recognition and naming. In particular, a clear distinction should be made between priming based on

semantic (or categorical)<sup>1</sup> and associative relationships. Since in many studies no clear distinction has been made between these two forms of priming, we first clarify the distinction between *semantic* and *associative* priming.

Priming is a memory effect in which the exposure to a stimulus (prime) influences the response to a subsequent stimulus (target). The particular relationship between the prime and the target defines the nature of the priming effect. In particular, one can distinguish an associative relation among prime and target from a purely semantic relation.

In psycholinguistics the distinction between semantic and associative priming has been acknowledged and has generated numerous debates. In this context, semantic relatedness reflects the similarity in meaning or the overlap in featural descriptions of two words (e.g. “turkey-goose”). On the other hand, associative relatedness reflects the probability that one word will call to mind a second word (e.g. “cat-dog”). Associative relations are assumed to reflect word use rather than word meaning. Whether semantic priming can be observed in absence of associative priming and vice versa is a matter of considerable debate in word recognition literature.

Some authors argue that a purely semantic relationship between prime and target can provide very little priming effect or no effect at all [144, 219] and associative relationships are the main cause of what is generally referred as semantic priming. Shelton and Martin [219] claim that “words that are very similar in meaning or sharing many features will not show automatic semantic priming if they are not also associated” (p. 1204). Hutchison [111], nevertheless, challenged this conclusion arguing that it is possible to obtain a “pure semantic” priming effect when great care is taken to select semantically related but unassociated stimuli. Indeed, “pure semantic” priming effects have been demonstrated by McRae and Boisvert [156] and by Perea and Rosa [180].

On the other hand, other studies have demonstrated “pure associative” priming in absence of semantic relationship between prime and target [108, 260] and a dissociation between the two forms of priming has been also demonstrated within the same study [66].

A similar debate about whether priming effects are in fact due to associative or categorical relationships between items in memory has produced a similar division in face recognition literature, but the issue has not been fully investigated. Moreover, the two forms of primings are often confounded in this literature and there is an ambiguity about the locus of the effect. This is due to the fact that in many studies the stimulus pairs are simultaneously related by categori-

---

<sup>1</sup>These two terms are often used interchangeably in priming literature. Therefore in the first part of this section we will not distinguish between them. A more clear discrimination between terms will be provided at the end of the section.

cal and associative relationships and the term “semantic priming” is often used to include both kinds of relationships. For instance, McNeill and Burton [155] define semantic priming as “the fact that processing of an item is faster if it is preceded by a closely associated item” (p.1142) and explicitly state “throughout the body of this study we do not distinguish between semantic and associative relations”.

Carson and Burton [42] discriminated between “semantic”, “categorical” and “associative” relationships. According to them, two people are related “associatively” if they are “routinely observed together” (e.g. Bill and Hillary Clinton); they are “categorically” related if they “share a particular personal information” (e.g. Stan Laurel and Buster Keaton because they are two comedians); finally, they are “semantically” related if one or both of the previous relationships hold. According to this classification, semantic priming does not distinguish between categorical and associative relations. Furthermore, it should be noted that the definition of categorical relatedness is quite ambiguous. Bill and Hillary Clinton, for example, share “particular personal information” (e.g. they are married but they are also politicians). It is not straightforward from the previous definition, if “being married” is the cause of the associative link (i.e. they are routinely observed together because they are married) or is part of the semantic relationship. It is also unclear why the authors define “semantic” a relationship based on co-occurrence.

In addition, the idea that a certain amount of semantic information must be shared between two associate faces in order to produce a priming effect is implicit in several models of face recognition.

For instance, in the Burton, Bruce and Johnston’s model of person recognition [38] a form of semantic mediation is implicit.

We will briefly describe this model because it is probably the most influential model that offers an account for priming effects in person recognition and it is the model that has inspired the majority of studies on this issue. We refer to section 3.2 for more details and the comparison with other models of face recognition and naming.

This model comprises three sets of units of processing: Face Recognition Units (FRUs), Person Identity Nodes (PINs) and Semantic Information Units (SIUs). The units are organized into pool such that the units within a pool are connected to each other with inhibitory links. The links between units belonging to different pools are excitatory.

For each face there is a single face recognition unit which becomes active by matching the perceptual input from a familiar face. If a match is made then activation spreads from the FRU to the corresponding person identity node (PIN). A PIN is a multimodal unit that receives inputs also from other

systems (e.g., a PIN is activated by read names, voice and so on). It represents the access point to semantic information stored in SIUs and signals familiarity. When a certain threshold is crossed, the face is recognized as familiar and the PIN leads the activation to the corresponding SIUs. In turn, the activation of a SIU above its threshold corresponds to the retrieval of the corresponding personal information encoded into it (e.g. occupation). Each SIU is connected to the PINs of other persons who share the same attribute. For example, the SIUs representing a certain occupation are connected to the PINs of known persons with that occupation. Since the links between a PIN and relevant SIUs are bi-directional and excitatory, when a familiar face is presented, activation can spread to the representations of other persons linked to the same SIUs. In this model, categorical information organizes the connections between PINs and the only way in which two or more PINs could be associated is by semantic mediation. The IAC model does not include any mechanism to allow direct associative relationships between PINs and cannot explain purely associative priming effects without recurring to categorical information.

The model predicts that priming should be observed between two person sharing a common category. That is, the presentation of the face of a known politician (e.g. Barck Obama) should influence the speed of responses to a subsequently presented target person sharing the same occupational category with the prime (e.g. Nicolas Sarkozy). Highly associated pairs, it is suggested, do not differ qualitatively from purely categorically related pairs but simply possesses more conjoint SIUs. However, evidence from priming experiments are mixed.

Probably the first study that used faces as stimuli in a priming experiment was conducted by Bruce [34]. In this study, some of the prime-target pairs were defined as “good predictors” of one another (i.e. close associated items); other pairs, still related but not associated, were considered “bad predictors” (i.e. semantically related items). The results shown a very similar facilitatory effect in the two conditions, providing the first evidence of semantic priming in face recognition. However, Bruce’s results were based on extremely small samples of stimuli (5 related and 5 unrelated pairs in a sequence of 60 faces).

Other evidence supporting the view that semantic memory for famous persons has a categorical structure come from a study by Brennen and Bruce [31]. The authors reported significant categorical priming with face stimuli when they used a double familiarity decision task, but only an associative effect when subjects were asked to perform a single familiarity decision.

Carson and Burton [42] also presented results in favor of the IAC model. In four experiments the authors shown that it was possible to boost semantic priming effects when multiple primes were presented before the target. Since se-



semantic priming effect was found having similar characteristics to the associative effect, the authors suggest that semantic priming behaves like a weak version of associative priming and it should not be considered as a different mechanism.

This idea seems supported by a recent study by Vitkovitch et al. [253]. Using a competitor priming paradigm (that investigates the effect of the presentation of a prime three trials before a target on error rate and response latencies) it was found a similar facilitatory effect both when the prime-target pairs were closely associated and when they were non associated category members.

However, the idea that associative and semantic priming can be explained by the same underlying mechanism, is challenged by a recent study conducted by Wiese and Schweinberger [257]. These authors used reaction times and event-related potentials (ERPs) to study the organization principles of person-specific semantic knowledge by explicitly comparing effects of categorical and associative priming. Reaction times shown significant priming effects in both conditions but the amplitude and the scalp distribution of the ERPs to the target were significantly different (i.e. more positive over the central and parietal areas in associated condition, more posterior for categorical priming), suggesting that associative and categorical priming are based on at least partially different mechanisms.

In addition, other studies that tried to isolate categorical effects from associative effects within the same experiment, failed to observe categorical priming of person recognition. Young, Flude, Hellowell and Ellis [271] tried to determine whether mere membership in a certain category (i.e. occupational category) is enough to produce priming or whether, instead, an associative relationship between prime and target is an essential factor. They found that inclusion within the same category is not enough to produce priming, whereas associative relatedness is a strong predictor of it.

Barry, Johnston and Scalan [9] reported significant associative priming effects in face familiarity decision and face naming, but not categorical priming with occupation as shared category (two British comedians who have performed as sketch duo primed each other, but unrelated comedians did not). Based on this pattern of results, the authors proposed an alternative model in which memory representations of famous persons are structured in *biographical idiosyncratically organized gnostic* (BIOG) units that contain personal information, such as “British comedian”, “politician”, “came to fame in the 1960s” and so on. These units become associated through common episodic events, that is the experience of co-occurrence of people. For instance, when the BIOG unit of Oliver Hardy is activated, activity flows on to the connected unit of Stan Lauren not because both are comic actors but because both appeared together in the same movies.

To investigate the nature of associative relationships in face processing,

Vladeanu et al. [254] explored the effects of co-occurrence and semantic relatedness in face priming, using a learning paradigm with artificial, computer generated faces. The results showed that an associative priming effect can be obtained solely by the co-occurrence of computer generated faces, which have no semantic background that could explain their association. On the other hand, a priming effect is shown when prime-target pairs are semantically associated but never co-occurred. However the effect in this second case is weaker than that produced by co-occurrence. The author concluded that semantic and associative priming are two different phenomena but in many studies they are intrinsically interlinked and both factors may have contributed to the priming effects.

Echoing this conclusion, we believe that one of the main reasons of this overlapping is that in face priming studies associated stimuli share also a lot of categorical information. Consider, for example, one of the associated prime-target pair used by [9]: John Lennon and Paul McCarty. They are both persons, singers, male, British as well closely associated members of the Beatles. It could be argued that they are indeed categorically related but also share a significant degree of co-occurrence. If we find a priming effect using this pair, it is difficult to separate the contribution of categorical relatedness from that of associative relatedness. It is likely that both forms of primings contribute to the overall effect reported. Indeed, those studies that shown both categorical and associative priming effects within the same study, often reported a weaker effect when prime and target were categorically related than when they were closely associated (see for example [42]).

This could be one of the reason because the definition of associative priming is more fuzzy in person recognition than in word recognition literature.

Moreover, in contrast to object priming studies, where associated prime-target pairs may belong to very different basic level categories (e.g. carrot-donkey, squirrel-nut, cheese-mouse), in all the face priming studies cited here, items from the same basic level category (e.g. Eros Ramazzotti and Michelle Hunziker are both person) or subordinate category (e.g. Angelina Jolie and Brad Pitt are both actors) were used as stimuli. In other words, the only associative connections that have been studied to investigate the organization of person-specific information are associative links between entities belonging to the same category.

This is in line with a general view that considers people as “special entities” whose semantic knowledge differs in structure from that of other objects. This idea is supported, for example, by the results reported in [9]. The authors found qualitative differences in semantic and associative priming of faces and objects. Objects were primed reliably by both associates and semantically related non-associates. In contrast, for faces there was a substantial priming effects

for associated but not for semantically related items. The authors suggested that semantic representations of objects are inter-connected by abstracted superordinate categories but that representations of people are interconnected by networks of inter-personal relatedness rather than by categories. Since the associative links, as proposed by Barry et al. [9], are “social” in nature, it seems obvious that they may interconnect only “social” entities (i.e. persons).

We argue that associative relationships may be established between entities belonging to different categories. A building, for example, may be strongly associated to the city where it is placed (e.g. Colosseum-Rome), an artwork to its author (e.g. The Pietà-Michelangelo) or a product to its brand (500-Fiat). In this study we will investigate these kinds of associative relationships. The advantage to extend the definition of associative connections across the category boundaries is that the semantic relatedness between the associated entities is kept to a minimum.

This considerations lead us to clarify our use of categorical and associative relationships between entities.

We define these relationships as follows:

1. Two entities are said to be “categorically” related if they share the same basic level or subordinate level category. For example, Rome and London are categorically related because they are both cities (or capitals), as are The Golden Gate Bridge and The Rialto Bridge because they are bridges or Barack Obama and Nicolas Sarkozy because they are politicians (or Presidents). For the purpose of this paper, we do not consider semantic relationships at the superordinate level of abstraction (e.g. the relation between me and my neighbour’s dog because we are living things).
2. Two entities are said to be “associatively” related if the first entity call to mind the second entity and/or vice versa. We share with other authors the view that the primary mechanism for associative relatedness is that the two entities are routinely experienced together in the contexts in which they appear (i.e. both real and informational contexts). For example, The Buckingham Palace and London are associatively related in our definition if the entity London is produced in response to the entity Buckingham Palace and/or vice versa. The association could be created, for example, by the fact that people experience many episodic events that involve both entities (e.g. every time they go to London they visit The Buckingham Palace). Otherwise, the association could be also induced by the fact that London is cited very often when information about Buckingham Palace is provided or vice versa, or by the fact that pictures of London represent The Buckingham Palace and so on.

Even though we acknowledge that a purely associative relation can be induced by repeated co-occurrence of two entities otherwise unrelated and become automatically registered in memory (as shown in [254]), we argue that associative connections are in fact more likely to be determined by co-occurrence of entities related by meaningful relations. In other words, an entity co-occurs with another entity for a certain reason (e.g. Michelle Obama may be associated to Barack Obama because she is married to him and this link favor the co-occurrence of the two entities). A similar view has been proposed in word priming literature (see for example [165]). Moss et al. proposed that some words that have been traditionally thought to be related by co-occurrence in fact have a functional relationship. For example wallet and purse do co-occur often, but they do so because the function of the purse is to contain the wallet. Our view is not functional in the sense proposed by Moss [165] (we do not claim that an entity is associated to another because the function of the first is to perform an action on the second or vice versa), but it is functional in the sense that there is a (binary) property which connects two entities and these two entities may co-occur and become associatively related because of this property. According to this view, entities do not co-occur by chance in real life but they co-occur because there is a particular relation that connects them. Co-occurrence in turn strengthens the initial connection. We note, however, that not all the binary properties which connect entities necessarily create associative relations. You can know that Barack Obama was born in Hawaii and in this case there is a binary connection between these two entities in your memory representation, but it could be that Hawaii never call to your mind Barack Obama or vice versa.

This observation raises the idea that binary relations must be reinforced to produce the associative connection. Co-occurrence has a straightforward effect to reinforce the binary connections between entities and to transform an initial connection between them into an associative link. The more two entities co-occur, the stronger will be the association between them. So, the more Michelle Obama and Barack Obama are observed together in everyday experience (e.g. they appear together in official ceremonies, their are both cited in the same news articles or television programs, their pictures are shown on the same magazines), the stronger will be the associative connection between them. It is not the simple fact of being married that creates the associative connection between these two entities, but it is likely that since they are married they frequently co-occur in common episodic events. If this happens the association between them will be created.

We note also that the properties which may promote the creation of associative connections between individual entities are different from those that connect entities to their categories (i.e. ISA-properties). The first are connections that create a direct link between two entities (from that derives the name “binary” that indicates an entity-to-entity relation). The second are connections between one entity and the categories which it belongs to. ISA-properties are the connections that mediate categorical relationships between entities (that are indirect relationships), whereas binary properties are the connections underlying associative relationships.

Since binary properties convey semantic information, we prefer to use the term “categorical” instead to “semantic” to refer to relationships between entities that are mediated by categories.

The major purpose of the present study is to contrast categorical and associative priming for individual entities from different categories (person, artwork, building and product) in a recognition task. To investigate the recognition processes we used a familiarity decision task. In this task people are asked to make a decision about the familiarity of a target entity. The target entity is preceded by a stimulus (a written word or a picture) that can be differently related (associatively or semantically) to the target or unrelated to it. Priming effects are measured in terms of reaction time and accuracy.

### 7.3 Objectives and Rationale of the Study

The final goal of this study was to provide a contribution to clarify the organizing principles of entity-specific semantic knowledge. We can summarize the rationale of the study as follows:

1. Many studies have investigated priming of face recognition and face naming, but only few have compared faces with other kinds of entities. The attempt to compare faces and objects, for example, was motivated by the idea that faces are “special entities” and they are likely to be processed differently than other objects. However in these studies the non-faces objects were not accessed at exemplar level but at basic level or subordinate level. More precisely, we investigated whether there are qualitative differences in semantic and associative priming of individual entities. To this purpose, we compared the priming effects for entities of the person category with those for entities of other three categories: artwork, building and product. Priming effects were investigated using a familiarity decision task.
2. Entities can be related to each other either by being close associates or members of the same semantic category. Additionally, an entity can be

related to another entity of the same category (e.g. a person can be related to another person) or to an entity which belongs to a different category (e.g. a person can be related to an organization). To the best of our knowledge, all the studies that have investigated associative priming effects at the unique level of identity (i.e. face priming studies) have used stimuli belonging to the same basic or subordinate level category. As we have discussed above, this approach makes difficult to discriminate effects due to categorical relatedness (or common semantic features) from effects due to associative relatedness. For this reason, in this study we investigated priming effects in associated pairs whose members belong to different categories (e.g. Rome-Colosseum). Since previous studies have demonstrated the importance of associative relationships between persons, only for this category we decided to compare the effects of associative priming within and across the person category. In this way, we explored how memory representations of individual entities are inter-linked with those of other individual entities of the same or different category and explore whether these connections are qualitative different.

3. In their study on the nature of associative priming, Vladeanu, Lewis and Ellis [254] reported a strong associative priming effect based on simple visual co-occurrences of computer generated faces in the absence of any semantic-specific knowledge. This study examined only within-domain relationships among faces, implicitly assuming that co-occurrences between people are more frequent within modality (e.g. the face of Angelina Jolie is more often seen with the face than with the written name of Brad Pitt). We hypothesize that other kinds of co-occurrence can contribute to the creation of associative links. In particular, for entities belonging to different categories, across domain co-occurrence (e.g., the picture of Mona Lisa and the name of Leonardo da Vinci) is likely to be as much frequent as within co-occurrence, or even more frequent especially in informational contexts. Therefore, in this experiment we chose a cross-modal design to control for potential explanations of the observed effects by simple visual co-occurrences or direct connections within a given pool of processing units (i.e. face recognition units).

To investigate the points described above we examined associative and categorical priming in a entity recognition task.

## 7.4 An Entity Recognition Experiment

The goal of this experiment was to explore how singular representations of individual entities (i.e. singular concepts) are organized in memory. In particular, we investigated how these representations are interlinked with those of other individual entities belonging to the same or different basic level category and whether exist qualitative differences in associative and semantic relationships of individual entities which belong to different categories.

To address this issues we examined associative and semantic priming effects in an entity familiarity decision task, comparing person recognition with object recognition when both processes involve individual exemplars of the category. Individual objects were famous exemplars selected from three categories: artwork, building and product. Exemplars of the person category were famous people belonging to four occupational categories: politicians, singers, sport persons and actors.

### 7.4.1 Pilot Study: stimulus selection

For our experiment a) highly associated prime-target stimulus pairs, b) non-associated pairs belonging to the same category and c) unrelated pairs were required. To identify the prime-target pairs for use in the following experiment, we conducted a pilot study. The aim of this study was to identify an initial set of associated pairs from which we generated the complete list of experimental stimuli for the four conditions. To create this set, we compiled a list of famous entity names (12 entities for each category) to be used in a free association task. In this task, each name on the list had to be rated on a 4-point scale according to its familiarity to the participant (1=unfamiliar, 2=rather unfamiliar, 3= rather familiar, 4= familiar). These fame ratings were collected to ensure that the entities selected were really familiar to participants. In addition participants were asked to write down as many entities as possible that came spontaneously to mind when they encounter a particular name. These spontaneously generated names were assumed to be associated to that particular name on the list. This means that we took an entity B to be associatively related to an entity A, if B was produced in response to A by the majority of participants. Fifteen participants took part in this pilot experiment (8 females, 7 male). For each entity on the list we calculated a) mean fame ratings b) the frequency of occurrence of each name associated to a specific entity name. The most highly associated pairs were identified and combined into prime-target pairs for use in the following experiment.

For each of the second member of these pairs of associates, an entity who was not associated but who was from the same basic level category as the first

was selected<sup>2</sup>. That is, we took an entity B to be categorically related to an entity A if both entities belong to the same basic level category (e.g. politician, painting, tower, car and so on) and B was not produced in response to A by any rater in the free association task.

Then, an unrelated but famous entity was chosen for each entity target.

The free association task produced some interesting patterns of results. First of all, for each entity on the list the prime-target pair with the highest degree of association was composed by entities belonging to different basic level categories.

For artworks, the most common association is with the author (75%) or with the place where they can be seen (25%). Buildings are more often associated with the place in which they are located, and more precisely with the city (92%). Finally, the most common association for a product is with its brand (100%).

The most heterogeneous pattern of results was found for entities of the person category. Persons are strongly associated to organizations (42%), artifacts (42%), places (12%).

Even though only in one case we found that the most frequent associated pair was composed by two persons, it should be noted that for seven person entities in the list we found that an associated person was mentioned by more than 50% of participants. This seems to confirm that associative relationships within the category are relevant for person entities. This result opens the question whether there are differences between associative connections within category and associative connections across category. In particular, the question is whether representations of people (i.e. singular concepts) are preferentially structured along social relationships, as hypothesized by Barry et al. [9] or these associative links are as strong as other associative connections with other types of entities. To address this issue, we added a subset of twelve associated pairs whose both members belonged to the person category.

These associates were selected with the same procedure described above, with the only difference that in this case a group of 15 judges were specifically asked to write down as many other names of famous persons as possible that came spontaneously to mind when they encountered the target person name. For each target member of these pairs of associates, a person who was not associated but who was from the same category as the target was selected by the authors. For the unrelated condition, a famous person of an unrelated category was chosen for each target.

The use of unrelated stimuli from the same category of the target gave us also the opportunity to verify whether there are differences between unrelated pairs whose members belong to the same high level category (i.e. person) but with different occupational role (e.g. an actor and a politician), and unrelated

---

<sup>2</sup>For the category person, pairs with the same occupational category were selected.



pairs whose members belong to different categories (e.g. person and artifact).

## 7.4.2 Method

### Participants

Eighteen participants took part in the experiment (11 female). Mean age was 30.61 years (SD=4.59) ranging from 23 to 40 years. Each participant was tested individually in a quiet room.

### Stimuli

For each category of the selected target entities (person, artwork, building and product), the experimental stimuli consisted of 12 pairs of closely associated famous entities, arranged into three sets (set A,B,C) of 4 pairs. In these sets the prime entities belonged to a different category than target entities. For the person category 12 pairs of closely associated famous entities from the same category (i.e. associated persons) were also used. In this way, we introduced a further condition, in which associated, categorical and unrelated primes were selected from the same category (i.e. Person). To distinguish between the two conditions in the person category, we named *Person Across* the condition in which target and associated primes were from different categories and *Person Within* the condition in which target and primes were from the same category.

Since in this study we used a cross-modality design, prime stimuli consisted of written names of entities, whereas targets were pictures (450×600 pixels in size) depicting the entities paired with the primes. Pictures were edited with Adobe Photoshop to remove background information (where present) and convert them to gray scale.

Each participant saw the entities in one set in their close-associate pairs, the entities in a second set rearranged to form pairs whose members were from the same category but no close associated, and in the remaining set rearranged to form unrelated pairs. The allocation of the sets to the experimental conditions was counterbalanced across participants. In addition, 60 unfamiliar entities (24 persons, 12 artifacts, 12 buildings and 12 products) were selected to serve as targets and combined with the same 60 primes to generate the familiarity decision demand. In this way, unfamiliar targets were also preceded by famous names and prime familiarity would have no predictive value for target familiarity. As a consequence of the adopted design, each prime was presented twice in the course of the experiment. We note that potential effects of prime repetition would have occurred in all experimental conditions in a comparable way and therefore cannot explain the differences between conditions. Appendix B.1 lists

the names of the entities, divided by categories, used in the experiment.

### Procedure

Subjects were tested individually in a quiet room. The experiment adopted the prime-target design typically used in semantic priming studies. In each trial, the prime (i.e. an entity name) was presented for 1000 ms followed by a fixation cross (200 ms) and the corresponding target picture. The target remained on the screen until the subjects made a manual yes/no response by pressing the “A” key or the “L” key. The two response keys were counterbalanced for hand across the participants. The experimental trials were preceded by eight practice trials. None of the items used in the practice trials were adopted in the experimental trials. Each trial was initiated by the response on the previous trial after an inter-trial interval of 1000 ms. In Figures 7.1 and 7.2, we report the design of two sample trials with associative and categorical primings respectively.

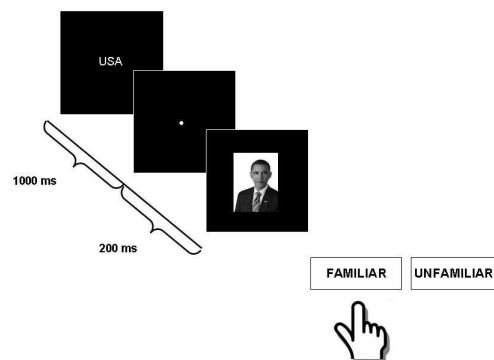


Figure 7.1: Trial presentation sequence in the entity recognition task (associative priming). On each trial, a word was viewed, followed by a picture, and the subjects were asked to indicate whether the entity depicted on the picture is a familiar entity or not.

Participants were instructed to respond only to the target picture. The task was to decide as fast as possible whether the entity depicted on the picture was a familiar entity or not. They were told that although they were not to respond to the name which preceded the picture, they were to pay attention to it as “in some trials it may help you to make your familiarity decision”. Response latency was taken as the delay between presentation of the stimulus target and initiation of a response as measured by the Matlab program. The presentation of the prime-target pairs was randomized by the computer separately for each subject. Each subject saw 120 experimental trials: 60 positive and 60 negative.

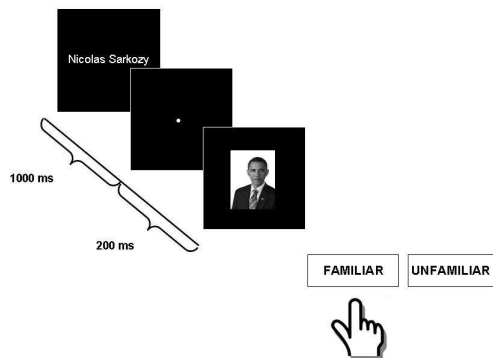


Figure 7.2: Trial presentation sequence in the entity recognition task (categorical priming). On each trial, a word was viewed, followed by a picture, and the subjects were asked to indicate whether the entity depicted on the picture is a familiar entity or not.

### 7.4.3 Results

The analysis was based on reaction times (RTs) of correct positive responses. Before the actual analysis, RTs from trials on which errors occurred were excluded from the analysis. Moreover, latencies over 2.5 s, which is equivalent to approximately 3 standard deviations from the mean ( $M_{RT} = 1116$  ms,  $SD_{RT} = 386$  ms) were discarded, as were outliers exceeding the participant mean by 2.5 standard deviations, for any particular condition.

Mean RTs for correct responses and accuracies are reported in Table 7.1.

Category	Measure	Primes		
		Associates	Same-category	Unrelated
Person Across	RT (SE)	937 (25)	1017 (21)	1082 (32)
Person Within	RT (SE)	919 (21)	1022 (33)	1045 (36)
Artwork	RT (SE)	954 (35)	1055 (42)	1052 (30)
Building	RT (SE)	998 (42)	1090 (46)	1098 (40)
Product	RT (SE)	961 (23)	1107 (44)	1148 (44)
Person Across	AC	0.95	0.98	0.90
Person Within	AC	1	0.88	0.94
Artwork	AC	1	0.91	0.94
Building	AC	0.90	0.97	0.92
Product	AC	0.97	0.95	0.85

Table 7.1: Mean Reaction Times (RT) in milliseconds (and Standard Errors (SE)) and Accuracies (AC) for Conditions of the Entity Recognition Experiment

For each category we performed a one-way repeated-measures analysis of

variance (ANOVA) calculated for mean RTs with prime type as a within-subjects factor (factor levels: associated, same-category, unrelated).

The analysis for the category Person Across resulted in a significant main effect,  $F(2, 34) = 11.39, p < 0.001$  (see Figure 7.3 for a graphical representation of mean response times for priming condition).

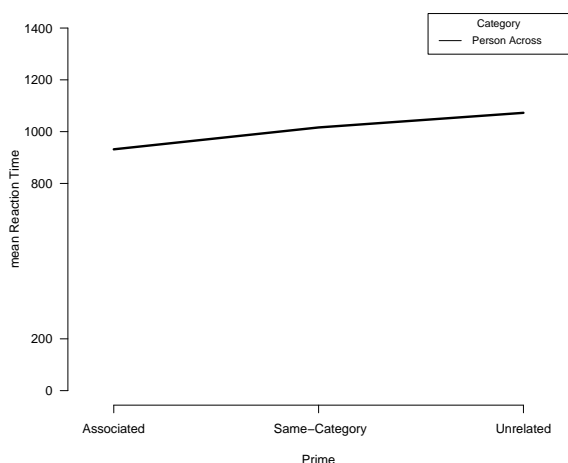


Figure 7.3: Mean response times for the category Person Across by prime condition.

Post hoc Tukey honestly significant difference (HSD) tests revealed significant faster RTs for the associated condition (937 ms) than for the same-category (1017 ms), ( $q = 4.02, p < 0.05$ ) and unrelated conditions (1082 ms), ( $q = 6.70, p < 0.001$ ). However, RTs in the same-category condition were not significantly different from those in the unrelated condition, ( $q = 2.68, p = 0.15$ ).

The same analysis for the Person Within category resulted in a significant main effect  $F(2, 34) = 9.03, p < 0.001$  (see Figure 7.4). The post hoc Tukey test revealed a significant difference between associated and same-category conditions, indicating faster responses for the associated condition (919 ms) than for the same-category condition (1022 ms) ( $q = 5.13, p < 0.01$ ). No significant difference was found between the same-category (1022 ms) and unrelated conditions (1045 ms) ( $q = 0.16, p = 0.99$ ).

A significant main effect (see figure 7.5) was also found for the category Artwork,  $F(2, 34) = 3.93, p < 0.05$ . The post hoc analysis showed faster responses for the associated condition (954 ms) than for the same-category condition (1055 ms), ( $q = 3.44, p < 0.05$ ). The comparison between the same-category (1055 ms) and the unrelated conditions (1052 ms) did not show significant difference.

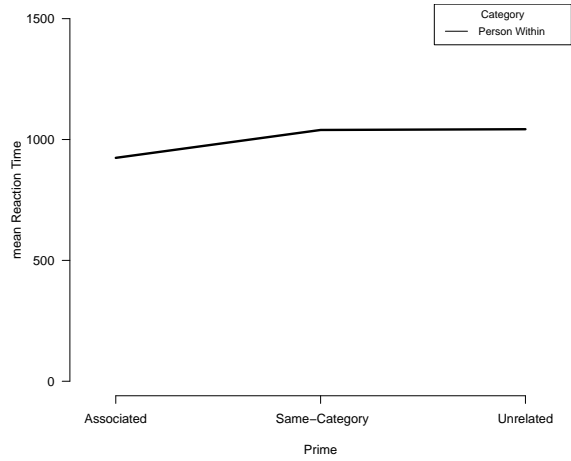


Figure 7.4: Mean response times for the Person Within category by prime condition.

( $q = 1.09$  ,  $p = 0.72$ ).

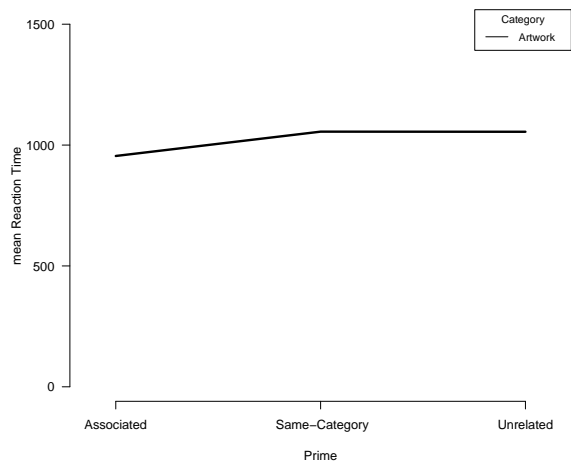


Figure 7.5: Mean response times for the Artwork category by prime condition.

As shown in Figure 7.6, the main effect was not significant for the Building category,  $F = 1.53$ ,  $p = 0.22$ , even though it shows a very similar trend than the other categories.

On the contrary, we found a significant main effect for the Product category

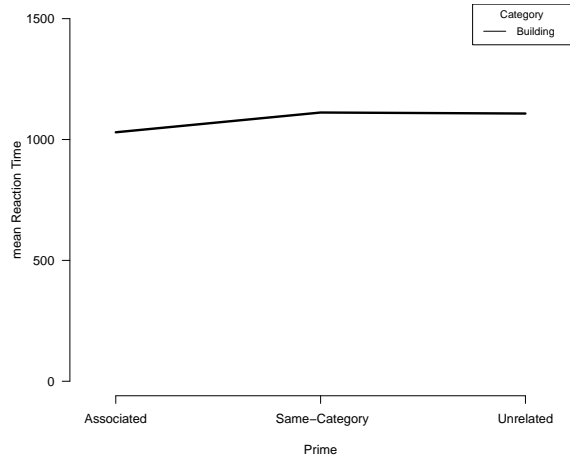


Figure 7.6: Mean response times for the Building category by prime condition.

$F(2, 34) = 5.63, p < 0.01$ . In Figure 7.7 the mean response times by prime condition for the Product category are shown. The post hoc analysis revealed the same pattern of results found for the other categories (with the exception of the Building category). In particular, we found a significant difference between associative and same-category conditions with responses that were faster in the associative condition (961 ms) than in the same-category condition (1107 ms), ( $q = 3.45, p < 0.05$ ), but no significant difference between same-category (1107 ms) and unrelated (1148 ms) conditions, ( $q = 1.09, p = 0.72$ ).

In order to test whether there are differences in semantic and associative priming for faces and objects, we collapsed person and object classes to form two general domains (i.e. Person and Object, respectively).

The comparison between the two general domains (face vs. object) was motivated as follows. First of all, we did not find a significant difference in the pattern of results obtained for the Across Person and the Within Person conditions (see Figure 7.8). Therefore the two person categories were collapsed to create a general Person category.

Second, we tested for differences between the three non-face categories (artwork, building and product). Mean reaction times were submitted to two-way ANOVA with category (artwork, building and product) and priming condition (associate, same category or unrelated) as within-participant factors. The main effect of priming condition was significant,  $F(2, 34) = 8.84, p < 0.001$ . Neither the main effect of category,  $F(2, 34) = 1.45, p = 0.24$ , nor the interaction was significant,  $F(4, 68) = 0.69, p = 0.60$ . The interaction plot for the three

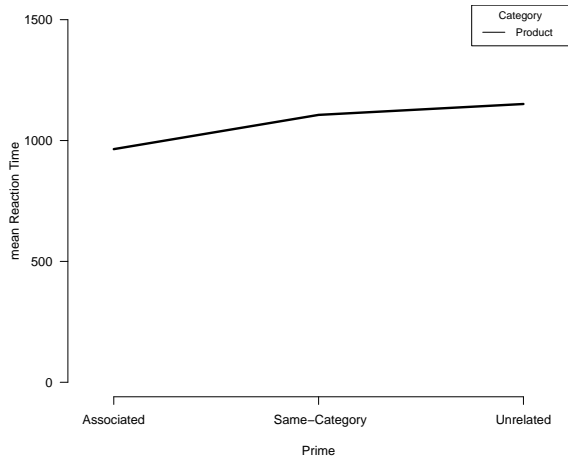


Figure 7.7: Mean response time for the Product category by prime condition.

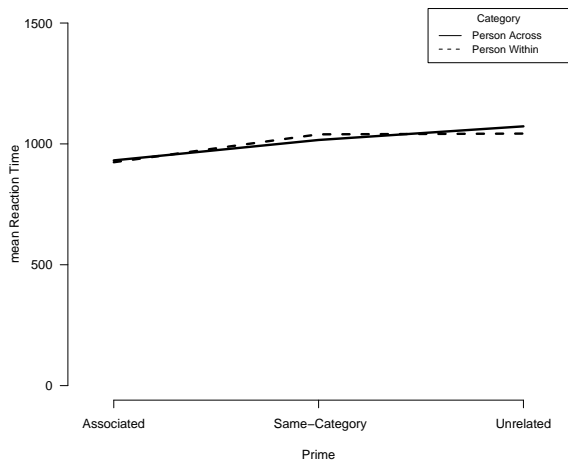


Figure 7.8: Mean response times for the Person Across and Person Within categories by prime condition.

categories and the three priming conditions is shown in Figure 7.9.

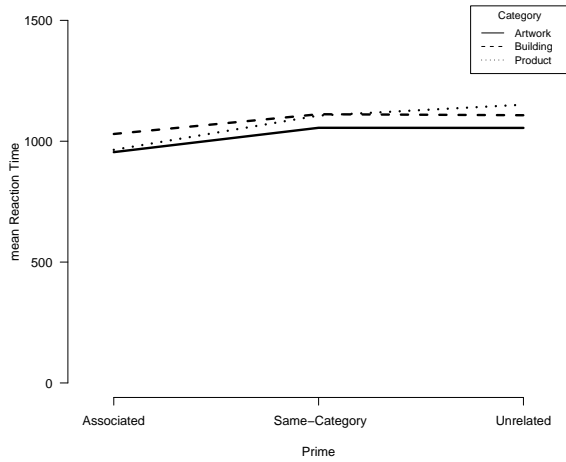


Figure 7.9: Mean response times for the three object categories by prime condition.

Due to the lack of the main effect for category and interaction, the three object categories were collapsed to obtain one individual mean RTs for the object-domain.

As a consequence of the aggregation procedure, we were able to compare semantic and categorical primings for faces and objects, by a two-way ( $2 \times 3$ ) analysis of variance (ANOVA) on the mean reaction times, with the variables of stimulus category (face vs. object) and priming condition (associate vs. same category vs. unrelated). The analysis was conducted by subjects with both variable (stimulus category and priming condition) as within subjects factors. The main effect of stimulus category was significant,  $F(1, 17) = 7.81, p < 0.05$ . A post hoc one-tailed t-test showed that responses to faces were significantly faster (1004 ms) than those to objects (1060 ms), ( $p < 0.05$ ). The main effect of priming condition was also significant,  $F(2, 34) = 22.92, p < 0.001$ . Post hoc Tukey honestly significant difference (HSD) tests revealed that: a) for faces the associate condition was significantly faster from both the same-category ( $q = 5.58, p < 0.01$ ) and the unrelated condition ( $q = 7.24, p < 0.001$ ), but the same-category condition did not differ significantly from the unrelated condition ( $q = 1.76, p = 0.47$ ); b) for objects the same pattern of results was found, that is the associate condition was significantly different from the same-category, ( $q = 4.37, p < 0.01$ ) and the unrelated condition, ( $q = 4.93, p < 0.01$ ), but the same-category condition did not differ significantly from the



unrelated condition, ( $q = 0.60$ ,  $p = 0.90$ ).

As shown in Figure 7.10, the stimulus category  $\times$  priming condition was not significant,  $F = 0.07$ ,  $p = 0.93$ .

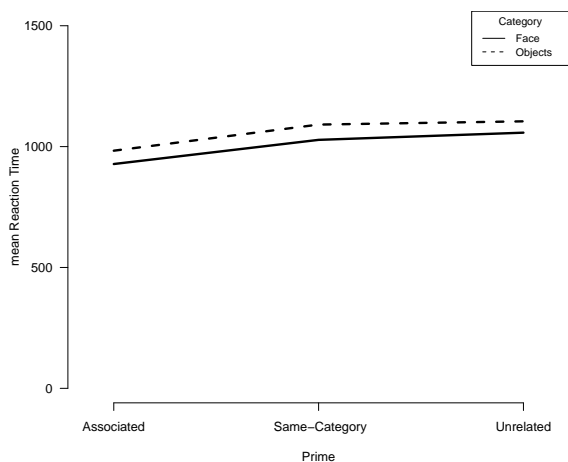


Figure 7.10: Mean reaction times for face recognition and object recognition at the three levels of priming condition

These results demonstrated that priming effects for faces were non significantly different (i.e. faster) from priming effects for objects. In particular, we found that for both categories, the associate condition was significantly different from the other two priming conditions, which did not differ significantly from each other.

## 7.5 Discussion

The entity recognition experiment produced a very clear and homogeneous pattern of results. For all the categories of entities used in the experiment, with the exception of the Building category, we found that entity familiarity decision times were reliably primed by the prior presentation of associates. On the contrary, non-associates from the same semantic category did not produce facilitation effects on familiarity decisions. This means that the time to recognize a familiar person, an artwork or a product was significantly and robustly facilitated by the prior presentation of the name of an associate entity, but was not reliably facilitated by the name of an entity from the same category but not associated.

Interestingly, the comparison between face and object categories did not

reveal a significant difference between the two domains in the amount of facilitation in the three conditions of priming. The only difference between the two domains was that the responses to faces were significantly faster than those to objects in all the priming conditions, confirming the astonishing ability of humans to recognize person identity from faces.

These results can be compared with those of Barry et al. [9]. The authors conducted two experiments which examined whether there exist differences in semantic and associative priming for faces and objects. Differently from our experiment, object stimuli used in their experiments represented generic objects (e.g. a table, a chair, a lion) which could not be recognized at the unique level of identity, but only at the basic level (e.g. as members of a general category). The authors found that faces were substantially primed by associates but not by non-associates of the same category. In contrast, they found that objects were primed reliably by both associates and categorically related non-associates. The results were interpreted as evidence for a different organization of the semantic knowledge of objects and people. We argue that to draw the conclusion that different processes underlie the organization and the access to semantic representation of faces and objects, a comparison between faces and objects at the same level of identity (i.e. as semantically unique entities) is required. To the best of our knowledge, our study is the first that performed this comparison in a priming experiment. Our results confirmed those of Barry et al. for faces, showing that face familiarity decisions were significantly facilitated by the prior presentation of associates from both the same and different category, but not by non-associated stimuli from the same category. However, contrary to Barry's et al. results, we found that object familiarity decisions presented the same priming effects than faces when the stimuli were recognized at the unique level of identity. These findings challenge the conclusion by Barry's et al. about a different organization of semantic representations of objects and faces and suggest a common mechanism to organize knowledge about individuals from different categories, as we will discuss in the next section.

Another important result is about the Person category. In our study we tested for differences between two different kinds of associative priming: priming across category (i.e. prime and target belonging to different categories) and priming within category (i.e. prime and target belonging to the same category). From our analysis we found that associated primes from different categories were as good as associated primes from the same category to produce priming facilitation. This result is important because previous studies which investigated associative priming effects for faces used associated primes from the same category (i.e. person), making difficult to isolate the associative effects from the categorical effects. In our study, we found that the associative effects in the two

conditions (within and across) were not distinguishable, producing evidence in favor of a pure associative facilitation. Therefore, it is clear that priming of entity familiarity decisions is associative but not reliably categorical and, at least for the Person category, both associated entities within the category and associated entities from different categories may facilitate the familiarity decision.

We note that in the Burton et al. model [38] of face recognition familiarity decision are proposed to be made by activating PINs which are proposed to be entry-level, threshold based recognition units that operate as amodal “gateways” to person information. According to this view, PINs would be activated by the recognition of names and voices as well as by FRUs. In this model a word prime would activate its PIN and corresponding SIUs. As there are proposed excitatory, bi-directional connections between PINs and SIUs, priming is interpreted in the terms of feedback activation from SIUs to increase the activation of PINs which are connected to the same SIUs. As we found priming effects from close associates but none from non-associated members of the same occupational category, then it would appear that only activation from the SIUs of associates feeds back to the PINs. Therefore, these results raise some questions about the nature of the elements of stored biographical knowledge and in particular whether it is correct to propose that these are represented by general categorical units (SIUs) such as “politician” or “actor” as proposed by the Burton et al. model. Moreover, the model can not explain the priming effects from associates belonging to different categories since SIUs are assumed to code only person-specific knowledge. On the contrary, our results are more compatible with a model in which singular representations of individuals from different categories can be connected directly through associative links, so that the activation of one of this singular concept spreads to all the associated singular concepts without the mediation of categorical units which are assumed to organize the knowledge of singular conceptual representations.

In our experiment we did not find a reliable priming facilitation for the Building category. Even though responses are globally faster for the condition with associated primes than for the condition with same-category primes, the difference did not reach the significance level, given the higher variability in responses for this category. In order to investigate possible differences between the stimuli used in the experiment and reveal possible sources of variability, we performed a post hoc analysis by items (paired t-test), comparing the associative and same-category conditions. We found a significant difference ( $p < 0.05$ ) in mean between the associative condition and the same-category condition for 7 of the 12 trials corresponding to the following associated pairs: Washington-White House, Bin Laden-Twin Towers, London-Big Ben, Berlin-Brandenburg Door, New York-Empire State Building, Paris-Louvre (see Table B.1 for the

corresponding same-category and unrelated pairs). The analysis suggested that associative priming effects can be observed for buildings, but that in some pairs used in the experiment the associative link may be not sufficiently strong to produce a priming effect. We noted that the strength of associate priming depends (among other things) on the frequency of co-occurrence. Specifically, the magnitude of priming would depend on the predictive value of the prime for the target. The predictive value is low when the frequency of co-occurrence is low but also when the co-occurrence is not specific (a prime co-occurs frequently with other targets). In our experiment the associative link between a building and its location or a person related to the building is less specific than other associative links such as the relation between an artwork and its author or a product with its brand. A place can be associated to many other things as well as buildings and this can explain the weaker association for some of the associated pairs tested in the experiment.

## 7.6 Implications for a Model of Entity Representation

The results of this experiment can be used to develop a model of the functional organization of semantic knowledge about individual entities which has at its core the notion of singular concept. More precisely, the model aims to explain how our semantic representations of individual entities, i.e. singular concepts, are inter-linked with those of other individual entities.

As described in section 7.2, there are two different ways in which individual entities can be related in memory. One is based on vertical relationships which connect individual instances to categories, the other is mediated by the horizontal relationships between individual instances within or across categories. We name “categorical” the relationships of the first type, “associative” the relationships of the second type (see Figure 7.11).

In the first case, abstracted superordinate categories are used to create a connection between individual items which belong to the same category. Two instances of the same category are connected to the representation of the category which they belong to and the category creates an indirect link between the two instances.

This means that once an instance of a category is presented and recognized, activation spreads to the other instances of the same category. If semantic representations of individual entities are inter-linked by categorical structures, we should register priming effects when prime and target entities has no other connection than the category membership.

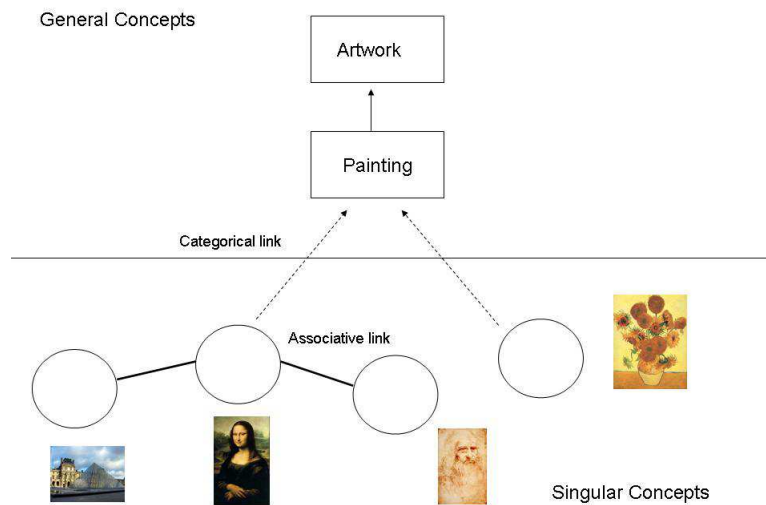


Figure 7.11: Associative and Categorical links between singular concepts

The second way by which semantic representations of individual entities can be structured and connected in memory is by means of direct associative links. These links are not mediated by shared category memberships but reflect meaningful co-occurrence relationships between singular representations of entities which non necessarily belong to the same category. Category membership can be part of the information shared by associated singular concepts but is not the semantic connection that interlinks them.

If associative links structure the representations of individual entities in memory, this means that once an entity is presented and recognized, activation spreads from the singular concept of the entity to its associated singular concepts. In terms of priming, we should obtain a priming effect when prime and target entities are associatively related even when they do not share category membership.

The priming effects obtained in the present experiment are more consistent with this second organization mechanism. Once an individual which can be recognized at the level of unique identity, such as a famous person, artwork, building or product, is presented and recognized, activation spreads to other individuals associatively connected to it, which produces the associative priming effects. The results of the experiment show clearly that there was no reliable categorical priming of individuals, in the sense that there was no significant benefit from primes corresponding to the proper names of members of the same category (e.g. another person from the same occupational category or another

painting) but not associated. Since we found that associative links between prime and target from different categories produced facilitation effects and for the category person we found similar priming effects when associated pairs were from the same or from different categories, it appears that activation within the semantic system spreads to the representations of associates by connecting paths other than those provided by general concepts.

These findings can be interpreted within a model that proposes that semantic representations of individuals (i.e. singular concepts) are inter-connected by networks of horizontal associative links rather than by vertical categorical relations. The singular concept of “Mona Lisa” is connected to the singular concept of “Leonardo da Vinci” by an associative link. Note that this link is associative not because the two entities are simply connected by the binary property “is created by” or the inverse property “is the creator of”. We argue that the information that connects the artwork with its author is part of the semantic information stored within the singular concept, as it is the information about the category membership. What it makes the binary property an associative link which organizes the connections between individual entities is the co-occurrence of the two entities in different contexts. Of course in this case the co-occurrence is due to the meaningful relationship between the artwork and its author, but the simple relationship is not enough to create an associative link. The associative link is created by experiential and episodic factors which reinforce the binary link. In this way, the activation of the Leonardo’s singular concept spreads to the singular concept of Mona Lisa, through the preferential route represented by the associative connection, producing a facilitation in recognizing the entity and retrieving information about that entity. On the contrary, the singular concept of “Mona Lisa” and that of “The Sunflowers” share the same category membership (i.e. both are paintings) but this shared membership does not inter-connect the two representations.

We do not deny that people use higher level categories to organize their knowledge about individuals. We can use category membership to connect “Mona Lisa” and “The Sunflowers” if we are required to list famous paintings, but this is not the main mechanism that structures our knowledge about individuals in memory. We propose that representations of known individuals are connected to each other individually by links representing specific associative relatedness. Contrary to the Barry et al.’s [9] model for face processing, we argue that horizontal links can be established not only within the person category as a consequence of “social” and “interpersonal” relationships (e.g. who is married to whom or who works with whom), but also between individuals from different categories which are connected by binary relationships reinforced by co-occurrence.

To conclude, we propose that singular concepts are organized within a network of horizontal associative links rather than being connected by vertical links with shared higher-level conceptual representations and this organization mechanism is not peculiar of singular concepts about people but it is the common way to connect singular concepts of individuals from different categories.





## Chapter 8

# Identification Relevance in Entity Representation

When we have introduced the notion of singular concept we have claimed that singular concepts can be represented as organized structures of semantic features (or attributes) which store our knowledge about the individuals they are about. But how is this information organized within a concept?

Many of the most influential theories of concepts and categorization used semantic features as their representational currency. For example, classical, prototype and exemplar theories of categorization all are based on featural representation [199, 159, 226], as are network models of semantic memory [49], connectionist models of semantics [183], vector models of memory [166] and similarity [249].

Many of these models do not assume that features may have graded relevance within a representation, but rather they assume the presence or absence of a feature (see for example [249]).

On the contrary, our assumption is that semantic features are of different importance in concept representation. Since the main function of a singular concept is the identity function - i.e. the function of providing the access to stored information that can be used to decide if an encountered entity corresponds to an entity previously encountered and stored in memory - it is quite natural to assume that the most important features in a singular concept are those which absolve better this function. This means that there are attributes that are more relevant than others to identify the unique individuals they represent. Intuitively, for example, the “name” of a person is more relevant to identify a person than her “occupation”, as well as her “eyes color” is more relevant than her “hair color”. One reason which can explain these differences

is that, for example, a person may change her occupation or hair color, but she unlikely may change her name or eye color.

We can consider the *identification relevance* of a feature as a measure of the contribution of the feature to the “identification core” of a singular concept. The “identification core” of a singular concept is thought to include those semantic features that enable to identify the referent of the concept (and to discriminate it from other similar referents).

Our notion of *identification relevance* is a variation of the notion of *semantic relevance* introduced by Sartori and Lombardi [207] to capture the importance of a given semantic feature in the distinction of one (general) concept from other similar ones. For example, the concept elephant may be more easily identified from the feature “has a trunk” than from the feature “has four legs”.

In this chapter we describe two studies which aimed to investigate the identification relevance of singular concepts belonging to five general types (i.e. person, organization, event, artifact and location).

The first study used a revised feature listing task paradigm to collect feature norms for singular concepts of entities from the five types reported above.

The second study used a more specific task, i.e. entity searching task, to explore which kinds of attributes people use to identify entities when they search information about them.

## 8.1 Semantic Feature Norms Production for Individual Entities

Many cognitive theories assume that semantic features are the building blocks of semantic representation (see for example [199, 159, 226, 49, 183, 166, 249]). Moreover, the attribute-value pair representation is the most often used knowledge representation scheme in information systems.

Given the importance of semantic features in shaping theories and representing knowledge, researchers have recently recognized the value of collecting empirically based semantic feature norms to construct conceptual representations that can be used for testing hypotheses, constructing experimental stimuli, and generating representations for implemented models [157].

These features norms have been used, for example, to derive measure of relevance for general concepts. Sartori and Lombardi [207] suggested that subjects’ verbal descriptions may be used to derive the relevant features of a concept. Going back to the previous example, the idea is that “has a trunk” is a semantic feature of high relevance for the concept elephant because most subjects use it to define elephant, whereas very few use the same feature to define other

concepts. “Has 4 legs”, on the other hand, is a semantic feature with lower relevance for the same concept because few subjects use it in the definition of elephant while using it in defining many other concepts.

In the same vein, we assume that people have conscious intuitions about the most important features for identifying individual entities and therefore they may be asked to derive, by description, relevant features for singular concepts.

The major goal of this study is to construct empirically derived entity representations in order to capture the features people consider most important to uniquely describe and identify individual entities, belonging to a few set of entity types. To this purpose we conducted a first study using a feature-listing task paradigm.

We argue that these data provide valid information about the cognitive representation of individual entities, not because they yield a literal record of semantic representations of entities, but rather because such representations are used systematically by participants when they have to generate entity descriptions. The basic premise of the method used is that participant’s conscious intuitions about the features relevant in singular concept representation actually map onto some underlying mental representation of the cognitive processing of singular concepts.

A participant’s list of attributes is assumed to represent a sort of temporary abstraction that contains the main attributes relevant for the identification (see for example [12] for a discussion about the dynamic realizations of concepts depending on context). These “online” representations are built in many entity-centric tasks (e.g., searching for information about entities). Therefore, the study has a second “technological” motivation. We argue that the basic information collected thorough this study can be relevant for the development of systems that manage information about entities (i.e., databases, ontologies, knowledge bases), as well for the development of entity-based methods for specific applications such as those required by an Entity Name System (see 5.3). This second motivation explains certain methodological choices which we followed in the study, such as the selection of the general categories of entities to be investigated which represented the first step of the research.

### **8.1.1 Method**

#### **Experimental Task**

In a typical feature-listing task, participants are presented with a set of category names and are asked to produce the attributes they think are important for each category. Since we were interested to collect norms for singular concepts, we needed to adapt the classical paradigm to our purposes. We considered two

different approaches that could be followed in our study.

The first is to define a small set of individuals from different categories, such as famous people, monuments or towns, and ask participants to produce the attributes they believe are important to identify those specific individuals. For example, we can present a picture or a written word of “Rome” (or both) and ask participants to list the features they think relevant for identifying it. A similar approach has been used by Gainotti et al. [75] in an experiment which aimed to evaluate whether subjective evaluations given by normal subjects confirm the different weight that various sources of knowledge have in representation of different biological and artifact categories and of unique entities. However in the experiment, the authors were interested to evaluate the influence of a limited and predefined set of sources of knowledge, such as perceptual knowledge (e.g. visual, auditory, tactile, olfactory and taste perceptions), motor and language-mediated encyclopedic information, but they did not investigate specific features. A limit of this approach to collect feature norms is that providing a small set of individuals for each category of entities under investigation may introduce a bias in generating features that can be tailored for the specific entities presented.

An alternative approach consists in inducing subjects to produce lists of attributes they think not generically important for a general category (e.g. person) but relevant to identify uniquely members of the category. According to this approach, people may be asked, for example to list the attributes they believe relevant to identify a specific individual which belongs to a given category (e.g. a specific person) without providing any specific exemplar. The advantage of this approach is that the descriptions produced by participants are not influenced by the selection of a predefined set of unique individuals and therefore should be more useful to identify a small set of features which are generally considered relevant for identifying the majority of the unique individuals of the category. In this study we adopted the latter approach.

It is worth to remark another important difference between a typical feature listing task and the task that we used in this study. In our version of the task we described a feature (or attribute) according to an attribute-value system. Each feature in an attribute-value system may possess a range of values. For example the attribute “color” may have different values, such as “red”, “blue”, “green” and so on. The features collected in a feature listing task are not attribute-value features but they simply are features which are present or absent. To make clear the distinction, people may use the feature “is used to cut” to describe a knife in a feature listing task. “Is used to cut” is a feature which contains its value and something may have this property (if is used to cut) or may have not the property (if it is not used to cut). On the contrary, the attribute “use” or

“function” is an attribute-value feature because it may have different values such as “is used to cut”, “is used to clean”, “is used to dry” and so on.

The typical feature listing task would require to present participants with a specific individual (e.g. Barack Obama) and ask them to produce features that are possessed by that specific individual (e.g. “is the USA President”).

Since different individuals may have different values for the same attribute and the typical paradigm would produce a different feature for each value (e.g. “lives in Italy”, “lives in Germany” and so on), we decided to force participants to list directly attribute types (e.g. Country). This approach was more convenient for the final analysis because we wanted to analyze patterns of attribute types and not specific attributes values for specific individuals.

### **Selection of Top Level Categories of Entities: an empirical motivation**

Defined the experimental task, the following step was to select an appropriate set of high-level categories for the experiment. Since one of the motivations of the study was to provide useful insights for the development of technological applications which manage knowledge about individual entities, we selected a small set of categories looking at the representation needs of a real application, i.e. the ENS described in 5.3.

The definition of “entity” in the ENS is purposely given in a very broad fashion, and covers all kinds of individual things from “anything that an information system talks about” to “an individual in an ontology” or “the interpretation of a variable in a first-order theory”. The reason for this very un-precise approach is the simple fact that – even though the creators of the idea have a sort of wishful thinking regarding the types of objects that should be covered – in reality it will be impossible to predict what finally enters into the system once it opens to the public.

The consequence is that in order to describe such entities in the ENS, it was decided to *not* impose or enforce a certain schema to be used for the description of different types of entities, as well as strong typing of entities is not pursued or enforced.

However, such genericity obviously has its downsides: the ENS can never *know* what type of entity it is dealing with, and how the entity is described, due to an absence of a formal model. This becomes very relevant when searching for an entity, a process called *entity matching*. The envisioned use of the ENS is that an agent (human or artificial) has a certain entity in mind and provides a description of this entity, which is then used for finding and re-using the entity identifier, similar to the use of a traditional search engine to find the desired target of an HTML hyperlink. However, the absence of information about the

type of the entity and its corresponding descriptions make difficult the use of specialized algorithms for entity marching.

To resolve this conflict between generality and precision, a possible solution would be to foster the convergence of entity descriptions on a small set of default types, and attributes for these types, by providing *suggestions*: when a new entity is to be created in the ENS, an agent has the possibility to select a default type and description, and “fill in the blanks”, or otherwise to provide any other kind of description. With this approach a useful clustering of entities could be achieved in the ENS, allowing the application of specialized matching algorithms.

Which is an appropriate collection of top-level categories which we can suggest to users for a “weak” or “light-weight” classification for the entities they create? This question provided the framework to define the categories of our study.

We identified four main requirements for this collection:

**Usefulness.** The set of top-level categories needs to be useful for a “normal” user, in that the concepts cannot be too abstract or too specific.

**Disjointness.** The categories need to be selected in a way that makes it easy to decide whether an entity belongs in one or the other, optimally through disjointness of the categories.

**Conciseness.** The number of categories should stay within easily manageable bounds, optimally below the “magic” number of 7 items [161], so that a user can decide at a single glance without further investigation which category should be chosen.

**Coverage.** The set of categories should be made in a way that all the entities that we envision to enter into the “population” of the ENS can be assigned to one of the categories.

In order to achieve these goals, we adopted a top-down approach: we analyzed the main top-level ontologies available in literature (Wordnet [78], Dolce [175, 149], Sumo [170] and Cyc [150]), to integrate important ontological distinctions from those ontologies. Even though the lack of correspondence between ontologies in their top-level division is well known representing one of the main obstacle to the integration of different ontologies, some important similarities can be identified (for a discussion in the context of ontology design see for example [173]). Starting from these similarities we tried to defined a set of few categories in accordance with our requirements.

At the end of our analysis we identified the following six top-level categories<sup>1</sup>:

- PERSON
- ORGANIZATION
- EVENT
- ARTIFACT
- LOCATION
- OTHER

We point out that the last category (OTHER) is a miscellaneous category that contains all entities that are not classifiable in one of the other categories and formally can be thought of as the complement of the union of the first five categories. Of course we did not include this category in our study.

Another aspect that should to be mentioned is the level of abstraction of our categories. Our choice was guided by two constraints. The first is related to the cognitive reliability of the categories. It is well-known that categories are organized into a hierarchy from the most general to the most specific, but the level that is cognitively most basic is in the *middle* of the hierarchy [198]. Many studies (see for example [160]) have shown that there is a preferential level of abstraction (named “basic level”) for object identification and description. At this level the categories are more differentiated and more attributes are reported to describe the members of the categories. This aspect is particularly relevant for our research since the experimental paradigm that we adopted asked people to identify different types of entities in terms of their relevant attributes. For this reason we needed to find a balance between too general or too specific categories.

The second constraint is more connected to the assumed use of the final system. Even though we do not know in advance the kinds of entities that will be entered in the system we can hypothesize that certain types of entities are more likely to populate the system than others.

These constraints explain for example the choice about the first category, PERSON. Although a more general category, such as Being, would allow us a better ontological coverage, including for example animals, it is not very probable that this latter type of entities would populate the system in large numbers. Moreover the level of abstraction of the Being category is quite far from the basic level making more difficult the task of listing attributes.

---

<sup>1</sup>We use small caps notation for the list because we want to denote the category itself, and not a natural-language label for the category. We could have chosen to use single characters as for variables, but decided to use this kind of notation of easier readability.

We also note that our categories have a good overlap with the main classes of names studied in the context of Named Entity Recognition (NER). Person, Organization, Location and Temporal Expressions are the most studied entity named types in the context of the NER task (see for example [94]) and recently there is a trend to include other entity types such as Product, Brand and Object [21] that correspond quite well to our Artifact category.

As evident from the list, we limited our analysis to a subclass of entities that we can describe as “physical” entities (things that have a position in space and/or time), missing out “abstract” entities (things that do not have spatial nor temporal qualities, and that are not qualities themselves). The distinction between physical and abstract entities is one of the most ubiquitous top-level division. It is at the base of the SUMO ontology (physical entity vs. abstract entity), the DOLCE ontology (endurant, perdurant particular vs. quality and abstract particular) and the CYC ontology (Intangible thing vs. Individual thing). The notion of abstraction is also present in WordNet, but has a different ontological coverage, not referring to state, psychological feature, action and phenomenon.

Following the distinction proposed by the CYC Ontology, we can distinguish between temporal entities and spatial entities, which justifies two of our top categories: EVENT and LOCATION. An EVENT is a thing that occupies a point (or period) in time, whereas a LOCATION is a thing that occupies a space. Both can have spatial and temporal parts, but the ontological nature is determined only by the *essential* parts that are temporal for events and spatial for locations.

Another important ontological assumption that we followed to build our list of top-level categories is related to the behavior of the entity in time. This distinction is connected to the difference between what philosophers usually call “continuants” and “occurrents”, or using the terminology adopted in the Dolce framework between “endurants” and “perdurants”. The main idea is that there are entities (endurants) that are wholly present (all their parts are present) at any time at which they exist and other entities (perdurants) that extend in time and are only partially present for any time at which they exist because some of their temporal parts may be not present. This motivated us to distinguish between entities that *are* in time like for example PERSON or ARTIFACT and entities that *happen* in time like EVENT, keeping another distinction that we can find both in the Sumo ontology (object vs process) and in the Dolce ontology (perdurant vs endurant).

A further ontological distinction we made within our basic categories is related to “agentivity”. This property refers to the attribution of intentions, desires and believes and the ability to act on those intentions, desires and believes. On the basis of this assumption we can distinguish physical entities that



are agentic such as PERSON (or groups of several agents operating together like ORGANIZATION), and entities that are not-agentic such as ARTIFACT.

Another difference that is taken into account is that between “Individual” entities and “Collection”. This ontological constrain is evident both in Sumo and CYC, and is used to explain the notion of collective entities such as ORGANIZATION, whose members can be added and subtracted without thereby changing the identity of the collective.

Similarly, WordNet distinguishes Entity (defined as something having concrete existence, living or non-living) and Group (which is any number of entities considered as a unit).

After making explicit the representation of the so-called ontological commitments (abstract vs physical, temporal vs spatial, endurant vs perdurant, agentic vs non-agentic, individual vs collective), we can provide definitions of each of our top-level categories (for a graphical representation see figure 8.1).

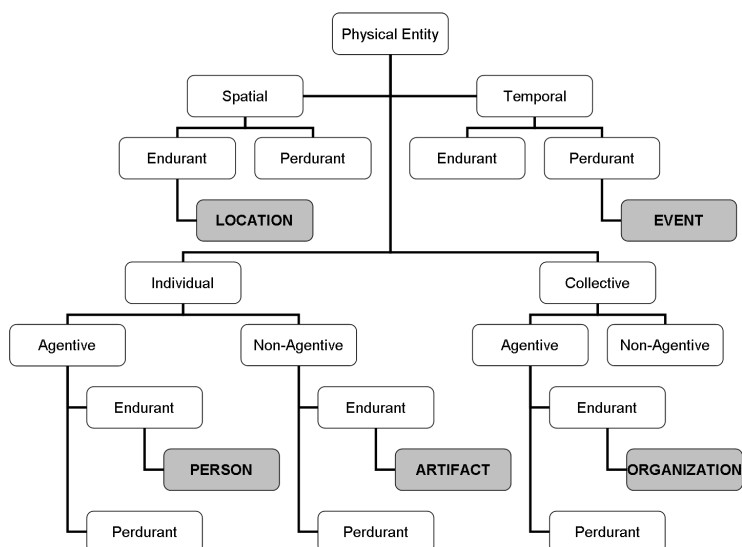


Figure 8.1: Top-level categories and Ontological Commitments

- *Person*: a physical entity, endowed with temporal parts that can change as a unit (endurant) and able to express desires, intentions and beliefs (agent).
- *Organization* : a physical collective entity, whose members are intelligent agents. In terms of behavior in time, an organization changes in time as a whole object so we can define it an endurant. As a collection of agents that operate together, an organization can be considered an agentic entity, characterized from desires, intentions and beliefs.

- *Event*: a physical individual entity that happens in time, perdurant.
- *Artifact*: a physical entity intentionally created by an agent (or a group of agents working together) to serve some purpose or perform some function. An artifact is a non-agentive endurant.
- *Location*: a physical individual entity that has a spatial extent, endurant.
- *Other*: any entity that cannot be categorized in any of the above categories.

## Tasks

Since our top level categories were at a high level of abstraction, we decided to introduce also a certain number of subcategories for each of them in addition to the top level category (named “neutral category”), reported in the section 8.1.1. This approach allowed us to investigate potential differences inside to the top level categories in terms of attributes reported, identifying (in addition of attributes common to all different subcategories) also possible specific attributes for specific subcategories.

Note that a similar approach has been recently proposed in Named Entity Recognition. In this context an increasing effort has been devoted to develop methods for automatically classifying entities into more fine-grained categorizations (see for example [68, 67]), exploiting differences in textual context rather than in attribute types.

In order to identify a small set of highly typical subcategories for each top level category, we performed a pretest asking eight people to list the most representative subcategories for each top level category. The categories were chosen on the basis of the frequency distribution of the answers, selecting the five most frequent subcategories for each category. We note that the idea was to define a subset of representative subcategories but our approach can be extended to other subcategories suggested by specific contexts.

For each top level category we developed 6 different scenarios one for each subcategory including the neutral category. We remark that for each top level category one scenario corresponded to the neutral category itself. This condition provided a way to compare the attributes common to all the subcategories of the same top level category with those reported for the top level category itself. By means of these scenarios we asked participants to imagine a specific entity from a given category (e.g. person) or subcategory (e.g. politician) and produce a list of all attributes relevant for uniquely identify that entity with the aim to obtain a unique profile of the entity. There was no restriction in the number

of attributes that could be reported. In table 8.1 we report the five lists of subcategories used in the experiment.

Scenario	Person	Organization	Event	Artifact	Location
1	politician	company	conference	product	tourist location
2	manager	association	meeting	artwork	city
3	professor	university	exhibition	building	shop
4	sports person	government	show	book	hotel
5	actor	agency	sport event	article of clothing	restaurant
6	person	organization	event	object	location

Table 8.1: Categories and Subcategories used in the experiment.

### Implementation

Each participant was randomly assigned to a combination of 5 scenarios, i.e. one scenario for each top level category. This was required to eliminate possible interference between different scenarios within the same top level category. To guarantee a balanced distribution of subjects to the different scenarios, we adopted a cyclic algorithm<sup>2</sup>. Through the first cycle the algorithm selected randomly one scenario from each of the 5 lists and assigned the combination of scenarios to the first subject. In the second cycle the algorithm selected the scenarios immediately subsequent (in order) to those assigned in the previous step. When all items of one list were assigned, the algorithm began again from the completed list. The order of the scenarios were randomized between participants.

The experiment was conducted in two different versions: English (eng), Italian (it) and was provided through the WWW. The subjects were invited (through email <sup>3</sup>) to participate in our online study. Once at this site, participants had to select the preferred language and were randomly assigned to an experimental condition, as described before; they then proceeded with 5 steps throughout the experiment: presentation, introduction, example, task and personal details.

Before starting the real task, participants were asked to read carefully the instructions which explained key terms used in the scenarios (for example the difference between “attributes” and “values” and the notion of “profile”). After that, a concrete example of the task was displayed. The domain of this example was deliberately chosen to be unrelated, to avoid that attributes reported as examples could interfere with the subsequent answers produced by subjects.

<sup>2</sup>The use of this procedure was necessary, because we could not counterbalance perfectly the assignment of participants to the different conditions given the online modality of the experiment.

<sup>3</sup>To spread the participation request we submitted our post to mailing lists such as DB-World or SIG-IRList

For the real task, the five scenarios were presented in succession (the order was randomized between subjects). Finally, a personal detail page was presented. The aim was collecting information about provenance, age, gender, internet experience and semantic web experience of participants to use for further analysis. This part of the experiment was optional and could be skipped.

As incentive to participation we arranged a lottery to assign a prize<sup>4</sup> among the participants who completed the task. Subjects were free to decide whether to participate in the lottery or not. In case of participation, they were asked to submit their email address, but the anonymity of the experiment was guaranteed by making sure that this information was not aggregated with the experimental data<sup>5</sup>.

### Subjects

We collected data from 353 participants (159 for the English version, 194 for the Italian version), 181 of these were male, 102 female, 70 did not report gender information.

The average age of participants was 31.06 years (SD=10.3),<sup>6</sup>. In table 8.2 we report the distribution of the number of subjects that specified their native country (262 out of 358), whereas in figure 8.2 we show the distribution in terms of Internet and Semantic Web experience, reported by 280 participants. From these self-evaluations it stands out that all subjects stated to have some knowledge in internet use and the majority of them reported “good” (117) or “expert”(134) knowledge. Differently, one-third of participants (102) reported none (54) or little knowledge (48) in the area of Semantic Web. Only 31 subjects defined themselves as experts in this area but a good part of participants reported “good” (85) or “average” (65) experience.

Country	N	Country	N
Italy	141	United Kingdom	5
Brazil	19	Netherlands	3
USA	14	Canada	3
Germany	14	Spain	3
India	11	Jordan	2
Pakistan	9	Malaysia	2
China	8	Mexico	2
Greece	6	Australia	2
Ireland	5	Switzerland	2
Others	21	<i>N<sub>tot</sub></i>	262

Table 8.2: Geographical provenance of participants.

<sup>4</sup>We gave away a medium-priced MP3 player.

<sup>5</sup>Every participant was represented in our database by a numerical id, with the intent of tracing the combination of scenarios, the corresponding answers and the anonymous personal details. The email address was stored disconnected from these records.

<sup>6</sup>considering only 285 subjects that actually provided age information

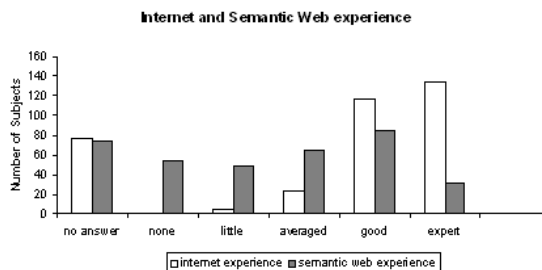


Figure 8.2: Self-evaluation of the participants regarding Internet and Semantic Web Experience

## 8.1.2 Results

### Normalization

As mentioned in the general description of the experiment, the peculiarity and the linguistic nature of the task made predictable a certain degree of variability in our data. To deal with this variability we normalized the data in three different steps: structural, morphological and semantic.

The first normalization step (structural) was performed mainly to report all answers in the form of lists of attributes. Indeed, although the instructions specified to insert one attribute per line in the specific form, some subjects disregarded this recommendation, using other break symbols (such as “,” “;,” “and” etc.) to separate the entries. Consequently, we had to implement a semi-automatic procedure to convert all the entries of our database in a standard form, splitting attributes so that the line number corresponded to the order of listing. This information will be extremely important for the future analysis on ranking. Moreover, in this first step, we checked the data to remove all typing errors.

The second normalization step (morphological) was finalized to report the attributes in a unique morphological form. For this purpose we removed articles, normalized the use of prepositions and the singular-plural inflections, we fixed the order for composed attributes (attributes which consists of two or more words).

Finally, the last normalization step (semantic), was conducted to aggregate attributes characterized by semantic overlaps (such as synonym expressions). In table 8.3 we report some examples of this preliminary phase. The number in brackets in the third column corresponds to the normalization step.

Attributes	Normalized form	Type of Normalization
name, address	name address	splitting (1)
surename	surname	typing error (1)
the name	name	article erasing (2)
date of birth	birth date	order (2)
near cities	neighbouring cities	semantic overlap (3)
zip code	post code	semantic overlap (3)

Table 8.3: Normalization examples. 1=structural normalization; 2= morphological normalization; 3=semantic normalization

## Measures and Results

For each category and subcategory, each feature was recorded with its production frequency, which is the number of participants who listed that feature for that concept. As we have seen discussing the semantic normalization procedure, a major issue in analyzing the features was to ensure that synonymous features were recorded identically, both within and among concepts. For example “occupation” and “profession” were considered synonyms. It was equally critical to ensure that features that differed in meaning were given distinct labels. To avoid potential ambiguity, responses were interpreted conservatively<sup>7</sup> and because of possible differences in meaning due to the language we started to analyze separately the data from the two linguistic versions of the experiment (Italian and English). The complete lists of features for the categories and the subcategories investigated in the study are reported in Appendix C.1. We describe the organization of the results for the category *Person* as example.

In table 8.1.2 we present the features listed by the participants for the category *Person*, for the Italian and English versions of the experiment, respectively. For each attribute we reported the absolute (F) and relative (f) frequencies<sup>8</sup>. Given the variety of the features reported by the participants (i.e. we obtained a long tail of unique features) we used a cutoff to include features, that is we considered only the features listed by at least 15% of the participants in each condition.

Since we found a good mapping between the attributes in the two language conditions we also performed an analysis on the aggregated data. In table 8.1.2 we report the results of the analysis on the aggregated data for the category *Person*. The results for the other categories are presented in Appendix C.1

<sup>7</sup>This means that we did not merge the features unless the overlapping was clear. For example, we decided to maintain separate the attribute “restaurant type” from the attribute “type of cuisine” even though it is plausible that a common way to classify restaurants is based on the type of cuisine. However, since other forms of classification underlying the general term “type” can not be excluded, we preferred to be as conservative as possible in the third phase of the normalization process.

<sup>8</sup>The relative frequency is the absolute frequency divided by N that is the number of subjects

Category	Attributes (it)	F	f	Attributes (eng)	F	f
<i>Politician</i>	age	17	0.56	party	24	0.76
	name	14	0.46	name	19	0.63
	political view	14	0.46	age	13	0.43
	party	13	0.43	country	10	0.33
	surname	11	0.43	gender	10	0.26
	type	11	0.36	role	8	0.26
	role	10	0.36	nationality	6	0.20
	education	9	0.30	surname	6	0.20
	experiences	7	0.23			
	curriculum	5	0.16			
		N=30		N=30		
<i>Manager</i>	name	13	0.46	name	0.71	16
	surname	11	0.39	age	0.28	7
	company	8	0.28	department	0.23	5
	age	7	0.25	experience	0.20	5
	role	7	0.21			
	type	6	0.21			
	education	6	0.21			
	N=28		N=21			
<i>Professor</i>	name	13	0.52	name	0.87	21
	specialization	16	0.64	university	0.41	10
	age	9	0.36	department	0.33	8
	surname	8	0.32	education	0.29	7
	educational institution	6	0.24	publication	0.29	7
	publications	5	0.20	age	0.20	5
	type	5	0.20	email	0.20	5
				research area	0.20	5
				surname	0.20	5
	N=25		N=24			
<i>Sportsperson</i>	type of sport	20	0.66	name	0.63	19
	age	14	0.46	type of sport	0.5	18
	name	14	0.46	age	0.33	10
	surname	9	0.23	gender	0.26	9
	type	7	0.23	birth-date	0.23	7
	birth date	6	0.20	nationality	0.16	5
	level	6	0.20	team	0.16	5
	N=30		N=26			
<i>Actor/actress</i>	age	16	0.51	name	0.88	16
	type	16	0.51	birth date	0.38	7
	name	15	0.48	movies	0.38	7
	experiences	14	0.45	gender	0.33	6
	nationality	11	0.35	country	0.27	5
	surname	10	0.32	age	0.22	4
	movies	10	0.32			
	birth date	7	0.22			
		N=31		N=18		
<i>Person</i> neutral category	name	20	0.74	name	0.73	19
	surname	17	0.62	gender	0.46	14
	birth-date	10	0.37	birth date	0.42	11
	age	10	0.37	age	0.38	10
	birth-place	8	0.37	education	0.23	6
	tax code	8	0.29	height	0.23	6
	occupation	7	0.29	nationality	0.23	6
	height	7	0.25	occupation	0.23	6
	place of residence	7	0.25	surname	0.23	6
	type	7	0.25	birth-place	0.19	5
	character	6	0.22	email	0.19	5
	weight	6	0.22	marital status	0.15	4
	eye color	5	0.18			
	nationality	5	0.18			
		N=27		N=26		

Table 8.4: Features and production frequencies for the category *Person*

Category	Attributes (all)	F	f	
<i>Politician</i>	party	37	0.61	
	name	33	0.55	
	age	30	0.50	
	role	23	0.38	
	experiences - career	19	0.31	
	political view	17	0.28	
	surname	17	0.28	
	education	13	0.21	
	country	11	0.18	
	type	11	0.18	
	gender	10	0.16	
		N=60		
	<i>Manager</i>	name	29	0.59
age		14	0.28	
role		12	0.24	
company		11	0.22	
experiences		10	0.20	
education		9	0.18	
competence		9	0.18	
		N=49		
<i>Professor</i>	name	34	0.69	
	specialization	20	0.40	
	age	14	0.28	
	surname	13	0.26	
	publications	12	0.24	
	university/ies	11	0.22	
	department	10	0.20	
		N=49		
<i>Sportsperson</i>	type of sport - specialty	38	0.63	
	name	33	0.55	
	age	24	0.40	
	birth date	13	0.21	
	gender	9	0.15	
	surname	9	0.15	
		N=60		
<i>Actor/actress</i>	name	31	0.63	
	age	20	0.40	
	type	18	0.36	
	movies	17	0.34	
	birth date	14	0.28	
	experiences	14	0.28	
	nationality	13	0.26	
	education	8	0.16	
		N=49		
<i>Person</i>	name	39	0.73	
	neutral category	23	0.43	
	birth date	21	0.39	
	age	20	0.37	
	birth place	15	0.28	
	gender	14	0.26	
	occupation	14	0.26	
	height	13	0.24	
	nationality	11	0.20	
	eyes color	8	0.15	
		N=53		

Table 8.5: Features and production frequencies for category *Person*: aggregated data (i.e. English and Italian).



To summarize the results it appears that a person is more likely identified by means of features about personal data (e.g. name, surname, birth date, birth place, age), followed by features about occupation and education (e.g. affiliation, specialization, competence, education, experiences) and finally by means of physical attributes (e.g. height, weight, eyes color). Dispositional attributes are less frequently reported by the participants.

An organization is most likely identified by means of its name, followed by features about the location (e.g. address, country), the kind of activity and objectives of the organization (e.g. business type, sector) and finally by aspects concerning its dimensions and internal structure (e.g. number of members or employees, turnover, faculties).

About events participants reported attributes concerning the location of the event and its temporal coordinates (e.g. date, time, duration), the topic and the title of it and finally they listed attributes about people involved in the event (e.g. participants, organizers, protagonists).

The features reported for the artifact category were more diversified. From the lists it comes out the predominance of perceptual features, such as color, material, shape and attributes about the dimensions of the artifact (e.g. size and weight). Other features for this category concerns the creator of the artifact (e.g. author, manufacturer, architect, artist). Finally artifact are identified by means of function or use.

A location is more likely identified by its geographical position or in relation to other locations. For example a city is identified by the country, an hotel, a restaurant or a shop by the city or the address. Other features can be used to specify qualitative aspects of the location (e.g. attractions, services, range of prices, number of stars) or quantitative aspects (dimensions, population, number of rooms).

### **Measuring the Identification Relevance of Features**

When we have introduced the motivations of this study we claimed that one goal was to provide useful results and measures for the development of systems which manage information about entities. Therefore, a second analysis of the data was performed with the aim to show a possible application of the results to a concrete representational issue in the context of the Entity Name System (ENS) described in 5.3.

As we have already mentioned, the ENS faces two problems: first, the system cannot assume to know what kind of entity it is dealing with (e.g. the entity of the user's request), and second, it cannot rely on homogeneous descriptions of entities (i.e. even if the type of the entity is known, it can not be assumed that

two entities of the same type are described using the same schema).

To resolve these problems a possible solution could be to foster the convergence of entity descriptions on a small set of default types, and attributes for these types, by providing suggestions: when a new entity is to be created in the ENS, an agent has the possibility to select a default type and description, and “fill in the blanks”, or otherwise to provide any other kind of description. With this approach the system can achieve useful clustering of entities, which put the system in a better position for entity matching, because at least in some cases it can understand better what kind of entity is described, and how it is described, which allows for a far better development and selection of specialized matching algorithms.

We argue that the data collected in our study can be used to establish the mentioned suggestions for entity types and their descriptions. Instead of simply accepting (or inventing out of mind) a certain schema, we propose to use a bottom-up approach of schema creation which exploit the results collected from a large sample of participants.

The data analysis was conducted having in mind two different issues: the first deals with the intent to provide a small set of default entity types suggesting a possible description through attributes, the second pertains the possibility of exploiting the information enclosed in the description provided by users to improve the efficacy of the entity matching algorithms. To such issues correspond two different questions: firstly, which is the information most frequently specified by subjects when they provide descriptions of entity types investigated? Secondly, which is the information more relevant to identify specific types of entities (distinguishing one type from others)?

To this purpose we adopted two measures: a measure of attribute dominance and a measure of attribute relevance. These measures can be considered a variation of those proposed by Sartori e Lombardi [207, 208, 209] in their model of Semantic Relevance for general concepts.

**Dominance** The problem of suggesting descriptions for types of entity at a high level of abstraction (corresponding to our top level categories) can be faced in two different ways.

The first consists of using directly the attributes reported for the scenarios of the neutral categories.

The second consists of identifying a set of general attributes used by subjects across the subcategories of the same top level category, aggregating the data of these subcategories (e.g. politician, manager, professor and so on for *Person*). The advantage of the second approach is twofold. First of all the analysis can be performed on a larger and diversified sample of observations. Secondly, the attributes shared by the subcategories represent an overlap that emerges from

the data (bottom-up) and is not the result of a high level abstraction operated by participants with the intent to provide a description of a generic entity classifiable into one of the top level entity types. In this sense, the descriptions of the subcategories are more close to the real descriptions that we expect in the final use of the ENS and can be useful to reveal some more details about the attributes that people need to describe real entities. However, if the selected subcategories are representative of the corresponding top level category, we should expect a substantial overlap between the results of the two kinds of analysis mentioned above<sup>9</sup>. The results of our study confirms this prediction as indicated by the asterisks in table 8.6 that mark the attributes that appear in the first 5 positions in both the analyses. In this section we present the measures that we adopted to perform the second kind of analysis.

When aggregating the data from the subcategories of each top level category, we require a measure to evaluate the importance of an attribute  $f$  for the top-level category  $c$ .

To this purpose we introduced a new measure that results from the combination of two components.

The first component is the *dominance* measure, that is a measure that quantifies the importance of an attribute for a specific category. We can formalize the function of Dominance ( $\phi: C \times F \rightarrow N$ ) in the following way:

$$\text{dominance} = \phi(c, f) = |\{s \in S : f \in F_s^c\}| \quad (8.1)$$

where  $S$  is the sample of subjects and  $F_s^c$  is the set of attributes listed by the subject  $s$  given the category  $c$ . In other words, the dominance  $\phi$  of the attribute  $f$  for the category  $c$  corresponds to the number of subjects that reported the attribute  $f$  for the category  $c$ . The dominance presents high scores when the attribute is frequently mentioned by subjects in identifying a member of the category.

Note that the dominance measure does not guarantee that attributes with high values of dominance are also attributes shared between the subcategories. If an attribute is reported by all participants for a specific subcategory (e.g., “political party” for *politician*) and only for this subcategory, it is possible that this attribute appears among the first dominant attributes for the corresponding top level category (e.g., *Person*), when the data are aggregated across the subcategories. For this reason the dominance measure is more suitable for the first kind of analysis that we have suggested before which is based on the data

---

<sup>9</sup>Note that the comparison of the two kinds of analysis provides an indirect method to test the representativeness of the selected subcategories. A lack of overlap would indicate that the subcategories are not representative of the category.

collected from the neutral categories that present a higher level of abstraction<sup>10</sup>.

Therefore, to derive a set of default attributes that are both frequently reported by subjects for a specific top level category, and also highly shared across the subcategories within the same top level category, we introduced a second component to our measure, *local sharedness*, that quantifies the level of sharing of an attribute  $f$  across a collection of subcategories:

$$sharedness_{loc} = \psi_l(f) = \frac{|Sc[f]|}{|Sc|} \quad (8.2)$$

where  $|Sc[f]|$  is the collection of the subcategories belonging to the category  $c$  that have in common the attribute  $f$ , and  $|Sc|$  is the collection of the subcategories of the category  $c$ .

Combining the two components listed above we obtained a new measure of dominance ( $\Psi$ ) that we name *local dominance*:

$$dominance_{loc} = \Psi(c, f) = \phi(c, f) * \psi_l(f) \quad (8.3)$$

Applying this measure to our data, we obtained the list of default attributes for our top-level categories. The first five attributes of our analysis for the two version of the experiment (English and Italian) are reported in Table 8.6. We note that the attribute more common across the categories is “name” which is the first attribute in two categories (Person and Organization) both in the Italian and in the English version and in the category Location but only in the English version. Moreover in the English version, “name” is present among the first 5 attributes in all the categories. Personal attributes (name, surname, age, gender, birth-date) are most frequently reported to describe people. In addition to “name”, organizations are identified in terms of spatial location (address, country) and type. Spatial (location) and time attributes (date, time) appear more relevant to describe events, whereas morphological and perceptual aspects (color, dimension, size, material) turn out to be more salient for the category Artifact. The most frequent attributes to describe locations are spatial (location, geographical coordinates, address, country).

**Relevance** The two measures of dominance described above do not provide information about the discriminatory power of an attribute  $f$  respect to a specific category  $c$ . If a user adopts a highly dominant attribute to describe an entity, we can not use this information to detect the presumptive category. The reason is that the dominance provide only a local evaluation of the importance of an attribute for the category without considering if the attribute is relevant

---

<sup>10</sup>We used the dominance measure to analyze the results of the neutral categories and to perform the comparison with the results of the second kind of analysis.

Category	English		Italian	
	Attributes	$\Psi(c, f)$	Attributes	$\Psi(c, f)$
<i>Person</i>	name*	110	name*	89
	age*	49	age*	73
	gender*	44	surname*	64
	birth-date*	29	type	56
	surname	24	birth-date*	34
		N= 145		N=171
<i>Organization</i>	name*	77	name*	87
	location *	37	type *	54
	country	34	objective/s*	44
	address	31	location*	37.5
	type *	23	head office	15.83
		N= 137		N=168
<i>Event</i>	location*	116	location*	126
	date*	69	date*	74
	time*	64	type*	68
	name*	49	time	57
	participants*	40	participants*	33.42
		N= 146		N=161
<i>Artifact</i>	color/s*	46	color/s*	74
	name*	33	type	60
	size*	29.16	dimension/s*	36
	type	28	material*	35
	price	20.83	price	28.33
		N= 140		N=168
<i>Location</i>	name*	86	location*	78
	country*	50	name*	73
	location	48	tipo (type)	57
	address	39.1	geographical position*	29.1
	geographical position*	35.83	address	18.66
		N= 145		N=169

Table 8.6: Local Dominance for selected top-level categories. We marked with an \* the attributes that appear in the first 5 positions of dominance also in the analysis which considered only the neutral categories.

also for others categories. Detecting those attribute which are dominant for a specific category but at the same time distinctive for it, is exactly the second aim of our research.

To identify attributes that correspond to this requirement, we propose a measure, named relevance ( $k$ ), that is the combination of two components: a local component (dominance) and a global component (distinctiveness). In the previous section we have formalized the first component. Now we pass to consider the second component.

The distinctiveness is a measure that quantify how much an attribute  $f$  is specific for a category  $c$ . When an attribute is used only in identifying one or few categories, its distinctiveness is high, whereas when it is used for many categories (or all) the distinctiveness score is low. The distinctiveness can be calculated as a function  $\psi_d(f) : F \rightarrow [0, 1]$  expressed as follows:

$$\text{distinctiveness} = \psi_d(f) = 1 - \psi_s(f) \quad (8.4)$$

where  $\psi_s(f)$  is a function of sharedness  $\psi_s(f) : F \rightarrow [0, 1]$

$$\text{sharedness} = \psi_s(f) = \frac{|C[f]|}{|C|} \quad (8.5)$$

where  $|C[f]|$  is the collection of the categories that have in common the attribute  $f$  and  $|C|$  is the collection of all categories. If an attribute  $f$  is listed for all categories  $\psi_d(f)$  is 0 and  $\psi_s(f)$  is 1.

The distinctiveness is a global measure because is transversal to all categories and in this sense it is category-independent and frequency-independent. This means that if we consider two different attributes  $f_1$  and  $f_2$ , one used by all subject only in the category  $c_1$  and the other used by only one subject only in the category  $c_2$ , their distinctiveness is identical ( $\psi_d(f_1) = \psi_d(f_2) = 1/|C|$ ) regardless of the category and the number of subjects.

We can combine the two measures (dominance and distinctiveness) in a single measure, the relevance  $k(c, f)$ , with the following formula:

$$k(c, f) = \phi(c, f) * \psi(f) \quad (8.6)$$

where  $\psi(f)$  is a logarithmic transformation of the distinctiveness  $\psi_d(f)$

$$\psi(f) = \ln \frac{|C|}{|C(f)|} \quad (8.7)$$

We point out that the idea to combine dominance and distinctiveness into a single measure of relevance has been adopted in other contexts. In information retrieval for example a similar measure (tf-idf) has been used to evaluate how

important a word is to a document in a corpus [206]. In cognitive science a model of semantic relevance has been recently proposed to compute the importance of a semantic feature in concept identification and has been used to explain semantic memory deficits (see for example [207, 209, 140]).

In our context we use this measure as an estimation of the contribution of an attribute  $f$  to identify an individual of a specific category  $c$ . We note that differently from distinctiveness, the relevance measure can be considered a concept-dependent measure. In other words, if the attribute is used by all (or the majority of) subjects to identify the category (high dominance) and is used only for that specific category (high distinctiveness), the relevance of the attribute for the category is consequently high. This means that the presence of that attribute is highly indicative (that is identifies with high probability) of the category considered. For example, the attribute “editor” is one of the most frequent attributes for the category *book* in both versions (it results in high values of dominance) and it is reported exclusively in the descriptions of that category (high values of distinctiveness). Combining dominance and distinctiveness, we obtain high values of relevance for this attribute when considered respect to the category *book*. Attributes with high values of relevance are highly informative for entity identification and entity matching. Continuing our example, consider the query  $q_1$ :<The Lord of the Rings and Allen & Unwin>. If we are able to recognize that “Allen & Unwin” is the name of a publisher, we can use this information for the entity identification and matching, because the presence of the attribute “publisher” suggests that the query refers most probably to the book rather than the movie that have the same title “The Lord of the Rings” (namely the same value for the attribute “title”).

In tables C.11, C.12, C.13, C.14 and C.15 we report the measures of relevance, considering the first 5 attributes for each subcategories. In general we can notice that in every subcategory stand out some highly specific attributes that combine high-middle value of dominance coupled with high level of distinctiveness. Just to make some example, “party” for the subcategory politician, “faculties” for university, “sport specialty” for sport event, “editor” for book or “number of stars” for hotel. In addition to these specific attributes, every subcategory presents two or three of those attributes that we identified at the top of the lists of dominance. These attributes are less distinctive for the particular subcategory (that is they are widely shared by the subcategories inside their top level category but are not extensively shared by other subcategories resulting in intermediate values of distinctiveness) but compensate with very high values of dominance. For example, “surname” and “age” are attributes of this kind for the category Person. A case apart is represented by the attribute “name”. As pointed above this attribute is the most shared between the subcategories

( $\psi_{it} = 0.93$ ,  $\psi_{eng} = 1$ ). However if we consider carefully the nature of this attribute we can note that the presumptive meaning of it could be very different in different contexts. For the category Person, “name” can mean “first name” or a combination of “first name” and “surname”<sup>11</sup>. For the category Company, “name” can be synonym of “brand” and legal constraints regulate the organization name assignment at least in local contexts. Normally, for products “name” is associated to a class of objects (i.e. iPhone 3G) with the same features and not to a single object (my iPhone). In the light of these differences, we decided to consider the attribute “name” distinct for the five top level categories. Using this expedient, we found that the attribute “name” appear nearly in all subcategories among the 5 most relevant attributes. In support of our methodological choice of aggregating data across the subcategories to obtain a list of general attributes as suggestions for entity description, we found that the most relevant attributes for the neutral categories correspond well enough to those found by means of the dominance measures obtained from aggregated data sets. The reason that why we adopted the aggregation strategy is primarily due to the size of the sample (the neutral category samples have about one sixth of subjects in comparison to the aggregated samples).

## Applications

As briefly sketched in the introduction of this study, the driving factors for this research were strongly related to its applicative potential. We present here two application areas concerning the functioning of the ENS: entity representation, and entity matching.

**1) Entity Representation** We can directly apply our findings to the way entities are represented in the Entity Name System in order to foster a certain convergence between how users *describe* entities, and how they *search* for entities.

Some of the client applications that are using the ENS today have been updated to give the user a selection of our top-level types, to manually classify an entity to be created. Subsequently, we provide the properties found to be most important for this entity type as a proposed “default schema” to the user, that can be manually filled with values.

As a second step, the knowledge we gained from investigating the co-occurrence of attributes enables us to work on a way to remove the manual classification step in favor of automatic classification. This is a more complex scenario that

---

<sup>11</sup>We suppose that the tendency of considering “name” as the combination of “first name” and “surname” is more likely for English speakers. Indeed in the Italian version of the experiment 63 participants (out of 89 that reported the attribute “name”) listed “name” and “surname” as two different attributes, whereas in the English version only 24 subjects (out of 110) listed the two attributes combined.



requires knowledge-based methods, which are mentioned among the ENS use cases. Imagine the description “Costa Forza Italia”: for a human (with some background knowledge), it is relatively easy to understand that we are describing a person called “Costa” who is member of the political party “Forza Italia”, and not – what would be another imaginable interpretation – a stretch of coast in Italy that is named “Forza”. The following steps facilitate such an automatic classification process:

1. Through the use of Named Entity Recognition (NER) functionality, which will be able to detect that “Forza Italia” is a political party; thus, we can tokenize the description into two parts:  $t_1 = \textit{Costa}$ , which is still unknown at this point, and  $t_2 = \textit{Forza Italia}$ , which we have just classified.
2. Relying on our findings, we can assume that “political party” is an attribute only relevant for politicians.
3. With the use of a background ontology (or a simpler structure that formalizes the results presented here), we can know that politicians are of type PERSON.
4. Based on our findings, we know that the most relevant attribute of PERSON is “name”, so we can argue that the token  $t_1$  is probably the name of the entity.

As a result, we can (a) provide a schema proposal to describe the entity, and (b) pre-populate the schema with the values already provided. We expect this to have significant positive influence on the “cleanliness” of data, and on the convergence between entity representation and entity matching, as we will explain in the following.

**2) Entity Matching** The second application area that we are directly interested in is entity matching, i.e. the attempt to return the single one entity that a user was (most probably) looking for when searching the ENS.

There are two ways how the research findings presented here can be applied to this problem: in a straight-forward manner, to take into serious account which descriptive attributes are more relevant for distinguishing entities, and a “backward” manner, by making inferences about the desired type of entity from a given search term.

The first case can be exploited by giving higher weights to the more relevant attribute types when ranking search results. To give a brief sketch, for example, as we have illustrated above, the “name” attribute usually has a high relevance; so for a search term  $x$  and two entities  $E_1 = \{\textit{name} = x\}$  and  $E_2 = \{\textit{place\_of\_birth} = x\}$ , it can be argued that  $E_1$  is the better match, because the search term appears in the more relevant feature.

A second way to make use of our findings is related to the issue of developing an advanced matching algorithm for a problem, by guessing the *type* of entity that is to be matched, based on co-occurrence of descriptive attributes. We are attempting to mimic human behaviour of “understanding” what is the intention behind a bag of search words, by applying the following steps:

1. First, we can perform automatic classification based on co-occurrence of attributes, similarly as explained before. The only difference is that now we are classifying a query string, to infer what kind of entity a user is searching for.
2. With the help of a thesaurus-based approach, we can approximate the “name” field in an entity description in different natural languages or representations (“nombre”, “nome”, “http://xmlns.com/foaf/0.1/name”, ...).
3. Finally, we can give assign an appropriately higher weight to this field when matching entities, as described before.

In the light of the result that “name” seems to be by far the most relevant attribute to describe entities, we do expect matching requests for entities to also reflect this phenomenon. We thus plan to directly apply the findings presented here to work on algorithms that work on co-occurrence of attributes similar to the example described above. Such algorithms will concentrate on (a) classifying what type of entity a matching request is most probably aiming at, and (b) relating search tokens to the most probable attributes of this entity type (i.e. which of the tokens most probably is the name of an entity, and which on is just “description”). To the best of our knowledge, this represents a novel approach, and we expect this to help us achieve higher-precision results without the a-priori knowledge (or enforcement) of any specific representational schema for entities. The insights presented here inspired a second study more focused on the search strategies used by people in formulating queries about individual entities.

## 8.2 An Entity Search Experiment

In the study described in 8.1 we adopted a feature listing task paradigm to investigate how people describe in terms of features individual entities from different categories and we proposed measures of relevance to quantify the importance that different features have for identifying these entities. The basic premise of the method was that participant’s conscious intuitions about the

most important features to identify individual entities actually reflect the underlying organization of the corresponding mental representations (i.e. singular concepts) in terms of feature relevance.

However, we note that the relevance of an attribute for identifying a given entity depends also on the specific context in which the identification process is performed. Our assumption is that, given a specific context, a person builds a sort of temporary representation of the entity which contains the most relevant attributes for the identification in that specific context. This view echoes the notion of flexibility and contextual-dependency in human (general) concepts proposed by Barsalou [10, 11]. In brief, the idea of the author is that there are two important types of properties associated with concepts, context-independent properties and context-dependent properties. Context-independent properties are activated on all occasions in which the concept is activated, whereas context-dependent properties are activated only by relevant contexts. In the same vein, we argue that some attributes are generally relevant to identify an entity (e.g. the title or the author of a book), while other attributes become relevant only in certain contexts (e.g. ISBN). If I talk about the last book of Umberto Eco with a friend of mine, it is implausible that I use the ISBN to identify the book during the conversation. However the ISBN can be used by a bookstore clerk to check for the edition of the book.

In this study we investigate how people identify individual entities in a very specific context: searching for information about individual entities by means of keyword queries.

Searching for information about individual entities such as persons, locations, events, is a major activity in Internet search. It was estimated, for instance, that searching for persons accounts for more than 5 percent of the current Web searches [95]. In a manual analysis on 1000 queries randomly selected from the search log of a commercial web search engine, Guo et al. [97] reported that named entities appear very frequently in queries and about 70% of the queries contain named entities.

General purpose search engines, like Google or Yahoo, are the most commonly used access point to entity-centric information. In this context, keyword queries are the primary means of retrieving information about a specific entity.

In this sense, queries for specific entities represent a variation of the expressed information need that has been studied in many Information Retrieval (IR) contexts [243, 233]. A query for a specific entity can be considered like a way to translate a human information need into a small number of attributes that the user considers relevant to identify the entity. Therefore, the analysis of real user queries should provide valuable insights into which kinds of attributes humans actually consider relevant to identify different types of entities during the search

process.

Several studies have looked at search engine log files to find out different aspects of the search process. These studies have revealed that the typical Web users only use a couple of query terms per query [109, 114], have short search sessions [223], typically check one result page and rarely use advanced query operators [13].

Besides general statistical analysis, a variety of other research topics have drawn interest (see Jansen and Pooch [113] and Spink and Jansen [231] for a review of the state of research in this field). Among the variety of research topics, we mention the reformulation process of queries over a period of time [191], query frequency distribution and caching [134, 264], search strategies and successful performance in Web searching [6].

Studies on what users search for are also reported in literature. On the one hand there are studies that focused mainly on term frequency distribution, co-occurrence of search terms and term clustering [202, 114, 263]. On the other hand, other studies have faced the problem to group queries into a small set of topics in order to produce a representation of users' search interests [232, 185] or to monitor the changes in popularity of topical interests [13]. However, from our review of the literature on query analysis it appears that little effort is made to investigate the semantic structure of textual queries.

Queries have been typically examined like list of terms (bag of words) without considering the semantic content of the keywords within the query. Our study aims to provide a contribution to this issue by investigating which attributes are considered more relevant by people to identify specific types of entities in a query formulation task and how these attributes are organized within the query.

As a first step towards a better understanding of this aspect of the query formulation process, we performed an experimental study. The goal of the study was to investigate the process that leads users to organize and represent their information needs using simple queries, limiting the analysis to queries that look for specific type of entities (Person, Organization, Event, Artifact and Location).

The motivation for this study was twofold.

First, from a cognitive point of view, the search task provided a real context where people were naturally forced to build on-line contextual-dependent representations of entities in which very few attributes (expressed by two or three keywords in a query) were used to uniquely identify these entities in the interaction with a search system. Exploring how people organize their information needs about unique entities in a search task provided a different way to investigate the identification relevance of attributes within a more realistic and constrained context, compared to that represented by the feature listing task.

Therefore, the results of this experiment could extend those obtained in the previous research about the internal structure of singular concepts. Moreover, the contextual need of selecting only few attributes to formulate a search query, would force people to use strategies tuned on the specific exigences of the task, individuating the attributes which are most relevant for the specific context of entity search. This aspect is strongly connected to the second motivation of our study which is related to applicative implications of the results.

From this perspective, understanding which attributes are considered more relevant by people to identify specific types of entities in a query formulation task could provide valuable insights for the development of information search systems. Therefore, the second motivation of the study, was to provide evidence for the beneficial impact that a cognitive study on the identification strategies used by people in entity search may have in improving the performance of computer-based search systems.

In particular, we aimed to show how our results could contribute to address one of the most critical problem in information retrieval that is the problem to capture the meaning of a query most likely intended by the user. Our assumption is that an important first step of performing such a task is to understand what type of entity the user is looking for. We call this process Entity Type Disambiguation. To address this problem we propose a Bayesian Model based on the assumption that an entity type can be inferred from the attributes a user specifies in a search query. Our aim was to apply the model to the queries collected in the experiment to test the performance of the model on the entity type disambiguation of real-world queries. Finally, to show the beneficial impact of the entity type disambiguation approach on a search system we aimed to test the effect of the disambiguation on the performance of a real system.

In summary, the study explores four main issues:

1. to investigate which attributes are considered more relevant by people to identify specific types of entities in a query formulation task;
2. to test the main assumption of a probabilistic (Bayesian) model for Entity Type Disambiguation that the entity type of the target of a query can be inferred by the specific pattern of attributes specified within the query;
3. to identify significant patterns of attributes that reproduce recurrent strategies in organizing the information in entity searching and show how these patterns can be integrated in the Entity Type Disambiguation model, improving the performance of it;
4. to provide evidence for the beneficial impact of Entity Type Disambiguation on the performance of a real search system.

## 8.2.1 Method

### Participants

301 participants took part in the experiment (165 male, 101 female, the others did not provide personal information). The average age of participants was 31.4 years (SD=9). Personal information (gender, age, country, Web and search engine experience) was collected through a questionnaire presented at the end of the experiment.<sup>12</sup> The participant's provenance is shown in Figure 8.3. In Web usage the participants' own evaluation revealed high experience with an average of 10.1 years of Web experience (SD=3.64) and an average of 5.87 (SD=3.7) hours of use per day. All subjects mentioned using Internet more than once per week with 233 participants (85%) using Internet daily, 32 (12%) almost daily and 5 twice per week (2%). In search engine usage the frequency of use was lower with 175 participants using search engine daily (65%), 71 almost daily (26%), 20 (8%) twice per week and 2 (1%) only once per week.

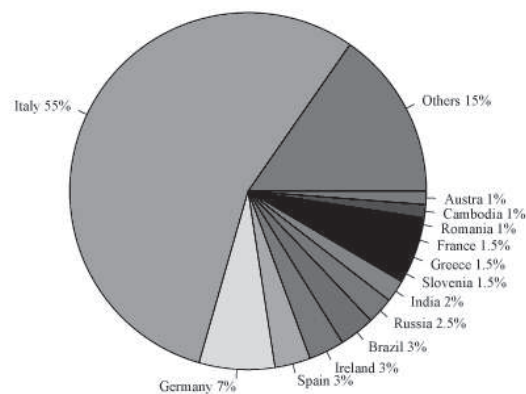


Figure 8.3: Geographical provenance of participants of the entity search experiment.

---

<sup>12</sup>Note that the questionnaire was optional, therefore the personal information statistics were calculated on the subset of participants which completed the questionnaire.

## Procedure

To answer our research questions we conducted a Web-based experiment with a significant amount of users (N=301). The advantage of the on-line modality is twofold. First, the target of our research is a user population that has experience with Web-based information retrieval systems and the Internet provides a “natural” environment to reach this target. Second, the Internet experiment allowed us to access a more diverse pool of participants (demographically and culturally), as can be noted in Figure 8.3, which is more representative of the real user population. For this reason, the experiment was performed in two versions, English (133 participants) and Italian (168 participants). Altogether these advantages contributed to improve the ecological validity of our research. However, the main drawback of the open environment of the Internet was the loss of some experimental control, such as providing supplementary clarifications about the task during the experimental session. To make sure that the instructions were clear enough and the interface was appropriate, the experiment was pilot-tested with two participants.

The experiment consisted of ten query formulation tasks. Participants were presented with an entity type (e.g., person) and they were asked to imagine any individual entity of their choosing belonging to this type (e.g., Barack Obama)<sup>13</sup>. Once the individual entity was chosen, participants were asked to formulate a query with the intent to find the homepage or an official Web site dedicated to the entity considered. In our example a plausible query may be <Barack Obama president USA>.

Every participant was asked to perform ten such tasks, submitting their queries through a dummy search engine interface (see Figure 8.4). Five tasks presented entity types at a very high level of abstraction. We call these types *high-level entity types* (person, organization, event, artifact and location). All the participants were tested on all the high-level classes. The other five tasks corresponded to more specific entity types (*low-level entity types*), selected from a predefined set of possible subtypes for each high-level type. Every participant performed only one low-level task for each high-level entity type. The task order was randomized between subjects. In the table 8.7 we report the complete list of high-level and corresponding low-level types. We note that high-level and low-level entity types were the same used in the experiment described in 8.1.

Using this experimental procedure, we collected a set of queries, for each of the entity types and subtypes, creating a suitable dataset for investigating our research questions. In particular, the selected experimental procedure which

---

<sup>13</sup>We remark that participants were provided only with information about the entity type, but they were free to choose any specific entity they came to their mind to perform the query formulation task.



Figure 8.4: Search interface used in the experiment to collect the participants' queries. In figure is shown the trial about a person search.

Person	Organization	Event	Artifact	Location
politician	company	conference	product	tourist location
manager	association	meeting	artwork	city
professor	university	exhibition	building	shop
sports person	government	show	book	hotel
actor	agency	sports event	article of clothing	restaurant

Table 8.7: Entity types and subtypes used in the entity search experiment.

allowed us to know in advance the intended entity type of each query collected by the participants, helped to create the annotated training set for the Bayesian Model which we used in the query analysis.

## 8.2.2 A Naive Bayes Model of Attribute Relevance

In order to address the first two issues of our study, we analyzed the data collected in the experiment within a Bayesian framework. Therefore, we perform the formulation of our problem, through a parallel introduction into the basic theory of the Naive Bayesian Model (NBM) used for the analysis.

We can represent a query  $Q$  as a set of unknown terms  $T = (t_1, t_2, \dots, t_n)$ , each of which can be a single word or a combination of words. We assume that each term  $t$  specifies the value of an attribute  $a$ . Assume that  $A = (a_1, a_2, \dots, a_n)$  is a set of attribute types. We map every term  $t$  into one appropriate type in  $A$ . After this mapping is established,  $Q$  can be represented by a vector  $\mathbf{a}$  (an assignment of attribute types  $a_1, a_2, \dots, a_s$  to the terms in  $T$ ). Finally, suppose that  $E = (e_1, e_2, \dots, e_m)$  is a small number of entity types.

The goal of our method is to define a Naive Bayesian model that can assign the most likely entity type  $e^*$  to a given query  $Q$  described by its attribute vector.

This is done by computing the probability  $P(E = e_k | A = \mathbf{a})$  for each possible entity type  $e_k$  and finally assigning  $Q$  to the type that achieves the highest



Query	Syntactic Preproc.	Semantic Preproc.
$Q_1 = \text{ISCW 2010 Shangai}$	$t_1 = \text{ISWC}$ $t_2 = \text{2010}$ $t_3 = \text{Shangai}$	$t_1 \Rightarrow \text{event name}$ $t_2 \Rightarrow \text{date:year}$ $t_3 \Rightarrow \text{location (city)}$
$Q_2 = \text{McCain Republican}$	$t_1 = \text{McCain}$ $t_2 = \text{Republican}$	$t_1 \Rightarrow \text{surname}$ $t_2 \Rightarrow \text{political party}$

Table 8.8: Two-step Preprocessing

posterior probability. Using the Bayes'rule we have:

$$p(E = e_k | A = \mathbf{a}) = \frac{p(e_k) * p(\mathbf{a}|e_k)}{\sum_{i=1}^m p(e_i) * p(\mathbf{a}|e_i)} \quad (8.8)$$

The critical quantity in Equation 8.8 is  $p(\mathbf{a}|e_k)$ . Since the NBM assumes that the conditional probabilities of attributes are statistically independent (that means that the value of a particular attribute is unrelated to the value of any other attribute), we can decompose the likelihood to a product of terms:

$$p(\mathbf{a}|e_k) = \prod_{j=1}^s p(a_j|e_k) \quad (8.9)$$

Because we are interested only in the most probable entity type, the NBM can be described by the function *disambiguate* ( $f : \mathbf{a} \rightarrow (E)$ ) that takes as argument a vector  $\mathbf{a}$  of attributes and returns the most likely entity type  $e^*$ . This function is defined as follows:

$$disambigaute(\mathbf{a}) = \arg \max_{e_k \in E} \frac{p(e_k) * p(\mathbf{a}|e_k)}{\sum_{i=1}^m p(e_i) * p(\mathbf{a}|e_i)} \quad (8.10)$$

### 8.2.3 Results

#### Preprocessing

Before applying the Bayesian Model to our data we performed two steps of preprocessing (see table 8.8 for examples). The first step, i.e. *syntactic preprocessing*, involved extracting the terms from the queries. A term can be a single word or a combination of words (i.e., a collocation). In this phase we also cleaned the dataset from unusual queries such as blank queries (empty), strings with only punctuation marks or senseless queries. Once the terms have been extracted from the queries, they were mapped into the attribute type set  $A$ . This mapping corresponded to the second step of preprocessing: *semantic preprocessing*. In Table 8.8 we report two examples of the two-step preprocessing.

The first step was conducted in a semiautomatic way (i.e., the deletion of

empty queries and a rough tokenization by segmenting the text at each space were performed automatically but the assignment of words to terms was performed manually), whereas the semantic preprocessing was performed entirely manually.

## Queries

In our experiment we collected an amount of 4017 queries. The average query length was 2.04 terms (mode=2 and median=2), which is in line with the results reported in literature (see for example [113]). Over 35% contained only one term and less than 3% of the queries contained five or more terms. Almost none of the queries utilized Boolean operators (over 99%). In only ten queries the operator AND was used, whereas the use of other operators was inexistent. The analysis of the word frequency distribution showed a very limited usage of articles, prepositions, and conjunctions. The only word without content that appeared in the first 30 most frequently used words was the preposition “of”.

## Bayesian Relevance of Attribute Types

The first goal of our research is to identify which kinds of attributes humans consider relevant to identify different types of entities during the search process. In order to address this problem, we used the Bayesian model described above to determine the relevance of an attribute  $a_s$  for a given entity type  $e_k$ . The relevance of an attribute for an entity type measures the importance of the attribute in the search of the type of the entity. In the NBM framework this corresponds to compute the posterior probability  $p(e_k|a_s)$ :

$$p(E = e_k|A = a_s) = \frac{p(e_k) * p(a_s|e_k)}{\sum_{i=1}^m p(e_i) * p(a_s|e_i)} \quad (8.11)$$

Assuming all entity types are equally probable (equal priors), the term  $p(e)$  is constant across the categories and can be ignored. Moreover, since the size of the training set is small, the relative frequency estimates of probabilities  $p(a_s|e_k)$ , will not be reasonable: if the attribute type never appears for a specific entity type in the data set, its relative frequency estimate will be zero. This means that the denominator in 8.11 will be nullified. Instead, we applied the Laplace law of succession [87] to estimate  $p(a_s|e_k)$ <sup>14</sup>. The estimate of the probability  $p(a_s|e_k)$  is given as:

$$p(a_s|e_k) = \frac{N_{ks} + 1}{N_k + 2} \quad (8.12)$$

---

<sup>14</sup>The Laplace law of succession has been used only to calculate the Bayesian relevance of attribute types but it was not adopted in the application of the model to the Entity Type Disambiguation problem.

where  $N_{ks}$  is the number of queries of type  $e_k$  in which the attribute  $a_s$  occurs and  $N_k$  is the total number of queries of type  $e_k$ .

We note that the rationale underlying the Bayesian measure of relevance is similar to that expressed by the measure of identification relevance proposed in 8.1. The relevance of an attribute for a given entity type is not only dependent on how frequently the attribute is used to identify entities of that type in the search process, but it is also dependent on how frequently the attribute is used to identify entities from other types. In other words, if the attribute is used to identify only entities of a given type, the presence of that attribute is highly indicative of the fact that the query is about an entity of that type. On the contrary, the relevance of the attribute is as lower as higher is the use of the attribute in queries about other entity types.

To make a concrete example, “city name” is the most frequently used attribute when a user looks for a city, but the same attribute is also used in queries about other entity types, such as queries about restaurants, shops, buildings, people and so on (see the tables reported in Appendix C.2). Therefore, the presence of this attribute is for sure relevant for the entity type City, but its relevance is mitigated (as expressed by the denominator in Equation 8.11) by the fact that the attribute is used in queries of other types.

In Table 8.9 we report the results of applying the Bayesian Model described in Equation 8.11 for the five high-level entity types addressed in our experiments. To clarify the semantic distinctions between the attribute types used in our classification, some examples are shown in Table 8.10. For each entity type we list the attributes with the highest probability values of relevance ( $p(e|a) \geq 0.15$ ). In Table 8.11 we report the same analysis for the low-level entity types of Person. The results for the other low-level entity types are reported in Appendix C.2.2.

From an overall analysis of the results it turns out that for the majority of high-level entity types “name” is the most relevant attribute used by people to identify the target of their request. This result confirms the centrality of proper names within the referential expressions (see for example [127]), but also the significance of mental names within singular concepts. However not all entities can be identified by means of a name. For example, pieces of clothing, sometimes meetings, or governments are entity types identified preferentially by means of other attributes. A particular case is represented by the entity type “product”. Our analysis shows that the majority of products are identified by the “model name” and not by the proper name of a specific entity. This result is interesting if it is considered in the light of the results of the experiments described in chapter 6 and 7 which showed that products are identified by model names as other entities are identified by proper names. This result reveals another

Entity Type ( $e$ )	Attribute type ( $a$ )	$p(e a)$
<i>Person</i>	first name	0.85
	surname	0.84
	occupation	0.89
	middle name	0.69
	pseudonym	0.33
	area of interest/activity	0.21
	nationality	0.20
	organization name	0.05
<i>Organization</i>	organization type	0.88
	organization name	0.73
	area of interest/activity	0.54
	location name	0.07
<i>Event</i>	event name	0.96
	event type	0.95
	date:month	0.83
	date:year	0.81
	date:day	0.75
	location name	0.20
	topic	0.17
<i>Artifact</i>	artifact type	0.98
	features	0.90
	model name	0.89
	artifact name	0.86
	historical period/epoch	0.56
	nationality	0.50
	organization name/brand	0.13
<i>Location</i>	location type	0.84
	location name	0.65

Table 8.9: Bayesian Relevance: top-level entity types

Query Terms	Attribute Types
<b>Person:</b>	
$T_1$ =Johann, Sebastian, Bach	$A_1$ =first name, middle name, surname
$T_2$ =Madonna, singer	$A_2$ =pseudonym, occupation
$T_3$ =Tim, Berners-Lee, semantic web	$A_3$ = first name, surname, area of interest/activity
<b>Organization:</b>	
$T_1$ =Greenpeace, environment	$A_1$ =organization name, activity
$T_2$ =Emergency, onlus	$A_2$ =organization name, organization type
<b>Event:</b>	
$T_1$ =ISWC, international conference, 2008	$A_1$ =event name, event type, date:year
<b>Artifact:</b>	
$T_1$ =Audi A4	$A_1$ = model name
$T_2$ =Mona Lisa, oil, portrait	$A_2$ = artifact name, features, artifact type
$T_3$ =Discobolus, Ancient Greek	$A_3$ = artifact name, historical period
<b>Location:</b>	
$T_1$ =Louvre, museum, Paris	$A_1$ = location name, location type, location name

Table 8.10: Attribute Types: examples

important aspect of the identification process: only a subset of entities are prototypically *namable* entities (e.g. person). Since users need also to identify non-namable things in their queries, the problem of Entity Type Disambiguation can not be entirely solved by the detection of the named entity in a query and the classification of it into predefined classes (an example of this approach can be

<b>Entity Type (<math>e</math>)</b>	<b>Attribute type (<math>a</math>)</b>	<b><math>p(e a)</math></b>
<i>Politician</i>	party	0.77
	location: country	0.56
	role	0.37
	related event	0.30
	nationality	0.28
	title	0.24
	surname	0.21
	first name	0.20
<i>Manager</i>	occupation	0.55
	affiliation	0.33
	role	0.29
	location: country	0.16
	location: city	0.19
	surname	0.16
	first name	0.15
<i>Professor</i>	location: city	0.57
	title	0.40
	affiliation	0.40
	occupation	0.27
	area of interest/activity	0.21
	surname	0.21
	first name	0.20
<i>Sportsperson</i>	area of interest/activity	0.62
	related event	0.51
	location: country	0.20
	surname	0.21
	first name	0.20
	nationality	0.15
<i>Actor</i>	movies/series	0.79
	role	0.30
	nationality	0.29
	first name	0.24
	surname	0.21

Table 8.11: Bayesian Relevance: PERSON

found in [97]). Given a query like “guitar Jimi Hendrix 1967”, the named entities are “Jimi Hendrix” and “1967”<sup>15</sup>, but the target entity of the query is an artifact (a guitar, precisely the first guitar burned by the guitarist on stage in 1967). The example shows that the simple classification of the named entities can be uneffective to detect the type of the target entity of the query and supports the idea that the disambiguation can be improved by including information from different kinds of attribute, such as “organization type” for organizations (e.g. non profit), temporal attributes for events (e.g. date), qualitative attributes (e.g. “color” or “material”) for artifacts. However, as we will discuss later, the difficulty of automatically disambiguating these kinds of attributes may challenge the possibility to adopt them in real applications. Another interesting aspect is the use of the attribute “location” (e.g. city, country, province) to identify entities from different types, such as persons, organizations and events, and of course location. However, queries about locations (e.g. city, restaurant, hotel) very often present more than one attribute concerning a location. Usually the first attribute specifies the name of the target of the query, while the second

<sup>15</sup>we restrict the word “named” to those entities for which one or many rigid designators, as defined by Kripke [128], stands for the referent.

specifies the location of the target. For example, in a query like “Venice Italy” the first location, i.e. Venice, is the target, while Italy is the spacial context where the target is placed.

### Entity Type Disambiguation

The second goal of our study was to test the hypothesis that the entity type of a query can be inferred from the pattern of attributes a user specifies in the query. To address this issue, we proposed a Naive Bayes Model (NBM) for entity type disambiguation. The model is described in Section 8.2.2. In order to test our hypothesis, we applied the NBM to our experimental data within the Weka framework [261]. We conducted the analysis using the subset of queries of high-level entity types (N=1350 queries) as learning set. The results of the stratified cross validation<sup>16</sup> performed on the learning set is reported in Table 8.12 that shows the confusion matrix on the learning set.

Classified as →	Person	Organization	Event	Artifact	Location
Person	259	1	5	5	0
Organization	0	268	0	1	1
Event	6	3	261	0	0
Artifact	6	2	0	262	0
Location	1	2	1	3	263

Table 8.12: Confusion matrix: learning set. Each row of the matrix represents the instances in a predicted class, while each column represents the predictions made by the model.

1313 of the 1350 queries were correctly classified (97.25%), corresponding to a mean absolute error of 0.023 and root mean squared error of 0.0981. In Table 8.13 we report the results in terms of precision, recall, F-measure and ROC area<sup>17</sup>.

In order to test the generalization performance of the NBM, we used two different test sets. The first test set (TSa) was created by randomly selecting 125 queries (25 for each of the five entity types) out of the set of experimental queries

<sup>16</sup>Cross-validation is a technique for assessing how well the model which has been learned from some training data is going to perform on future as-yet-unseen data. The first step of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

<sup>17</sup>*Precision* is defined as the number of queries correctly assigned to the entity type divided by the total number of queries assigned to that type; *recall* is defined as the number of queries correctly assigned to the entity type divided by the total number of queries which should have been assigned to it; *F-measure* is the harmonic mean of precision and recall. Receiver Operating Characteristic (ROC) area is the area under the ROC curve that is a statistical technique using linear regression to describe the accuracy of the model by plotting predicted true positive rates (y-axis) at given false positive rates (x-axis). The larger the area under the curve the more accurate the model.

Class	Precision	Recall	F-Measure	ROC Area
Person	0.95	0.96	0.96	0.99
Organization	0.97	0.99	0.98	0.99
Event	0.98	0.97	0.97	0.99
Artifact	0.97	0.97	0.97	0.99
Location	0.99	0.97	0.98	0.99

Table 8.13: Learning set evaluation. TP=true positive; FP=false positive

about the low-level entity types<sup>18</sup>. The second test set (TSb) was created by extracting 125 queries for specific entities from a collection of queries provided by [1] to evaluate entity search and entity linkage methods. The aim of using this second test set was to evaluate whether the performance of the NBM on queries obtained from real applications (e.g., Wikipedia’s search system) was as good as that obtained for experimental queries. The results of this comparison are presented in 8.14.

Class	Precision		Recall		F-measure		ROC Area	
	TSa	TSb	TSa	TSb	TSa	TSb	TSa	TSb
Person	1	0.81	0.88	0.87	0.93	0.84	0.98	0.98
Organization	0.95	0.83	0.87	0.83	0.91	0.83	0.98	0.97
Event	0.91	0.95	0.87	0.87	0.89	0.91	0.96	0.97
Artifact	0.92	1	1	0.78	0.96	0.88	0.99	0.96
Location	0.85	0.76	1	0.92	0.92	0.83	0.99	0.97

Table 8.14: Test set evaluation

Overall, the performance of the NBM is very high and encouraging and the disambiguation model seems to perform well not only on queries collected in a controlled experimental setting but also on queries submitted to real search systems.

### Distribution Trends: Attribute Position

The third research question of our study was about the distribution of attributes inside the queries. We aimed to highlight possible trends of attributes that recur during the formulation process and that reflect, it is argued, the strategies used by users to organize their information need. To this purpose, we focused on the distribution of attributes in terms of position. If we represent a query  $Q$  as a vector of attribute types,  $\mathbf{a} = a_1, a_2, \dots, a_n$ , the position of an attribute type corresponds to the position of the corresponding element in the vector. The aim was to explore whether there is a preferential order followed by subjects when they organize the attributes within the query so that an attribute type is more likely used in a specific position in the query. For example, is the name of the target entity always the first attribute specified? In this case

<sup>18</sup>The 125 queries constituting our test set were not part of the sample which was used to run the learning phase.

the position of the attribute becomes extremely informative to understand the entity search process and should be included in an integrated model of attribute relevance. Consider, for example, the following two queries:  $Q1 = \text{“Silvio Berlusconi Mediaset”}$  and  $Q2 = \text{“Mediaset Silvio Berlusconi”}$ . The two queries contain exactly the same terms and consequently the same attribute types,  $Q1 = \text{“first name, surname, organization name”}$  and  $Q2 = \text{“organization name, first name, surname”}$ , respectively. The only difference between  $Q1$  and  $Q2$  is the order of their terms. For example, in  $Q1$  the attribute “first name” is in first position, whereas in  $Q2$  the same attribute is in second position, and so on. But do the two queries refer to the same entity target? or, reformulating the question in terms of entity types, is the entity type of  $Q1$  the same than the entity type of  $Q2$ ? If we submit the two queries to one of the most popular search engine, i.e. Google, and we look to the first results returned by the system we find that the two queries produce exactly the same results (see Figure 8.5 and 8.6). Our research question deals with exploring whether the two queries are equivalent from a cognitive point of view. This means to investigate if the order used to organize the terms (and therefore the attribute types) within a query conveys some information about the intended meaning (i.e. the target entity) underlying the query.

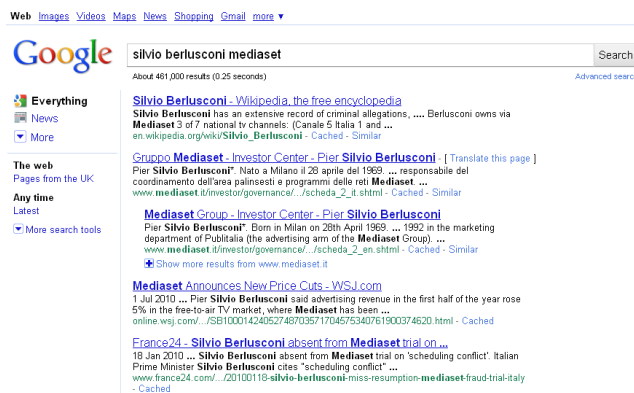


Figure 8.5: First five Google results for the query  $Q1 = \text{“Silvio Berlusconi Mediaset”}$ .

The Bayesian Model described in 8.2.2 does not make any assumption about the order of attribute types within the query, that is the probability of an entity type  $e_k$ , given an attribute type  $a_s$ , is the same independently from the position of  $a_s$  within the query. In our example, if  $A1$  is the vector of attribute types of  $Q1$  and  $A2$  is the vector of attribute types of  $Q2$ , the model predicts that  $p(E = e_k | A1) = p(E = e_k | A2)$  for each entity type  $e_k$ . For instance, the model predicts that the probability that  $Q1$  is about a person is the same than the



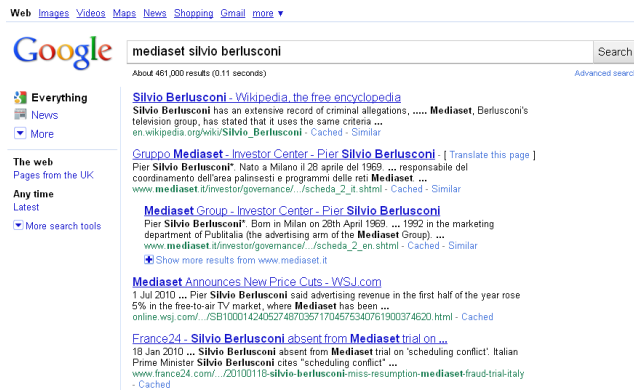


Figure 8.6: First five Google results for the query  $Q2=$ Mediaset Silvio Berlusconi.

probability that  $Q2$  is also about a person.

In order to investigate whether the attribute type position provides useful information about the intended target of a query, we analyzed the probability distribution of attributes by position within the query. To this purpose, for each attribute type  $a_s$  and entity type  $e_k$ , we calculated the probability of observing  $a_s$  in a given position  $j$ , as follows:

$$Pos_j = \frac{Npos_j}{N} \quad (8.13)$$

where  $Npos_j$  is the number of queries of type  $e_k$  in which the attribute  $a_s$  occurs in position  $j$  and  $N$  is the number of queries of type  $e_k$  in which the attribute  $a_s$  occurs, regardless of the position in the query. The analysis was conducted only on the queries about the five high-level entity types. A graphical representation of the results for the Person and Organization entity types is shown in Figure 8.7 and 8.8, respectively (see Appendix C.2.3 for the same analysis on the other entity types).

As shown in Figures the results of the position analysis give support to the initial hypothesis. Different attribute types present a preferential position within the query and at least the first two positions are significantly dominated by one attribute. For instance, we note that “first name” and “organization name” are the attributes with the highest probability in first position, respectively for Person and Organization. Instead, “surname” and “middle name” (for Person) and “organization type” and “activity” (for Organization) are the preferred attributes in second position. The analysis provides an interesting insight about the problem of the presumptive “cognitive equivalence” of the two queries,  $Q1=$ “Silvio Berlusconi Mediaset” and  $Q2=$ “Mediaset Silvio Berlusconi”, sug-

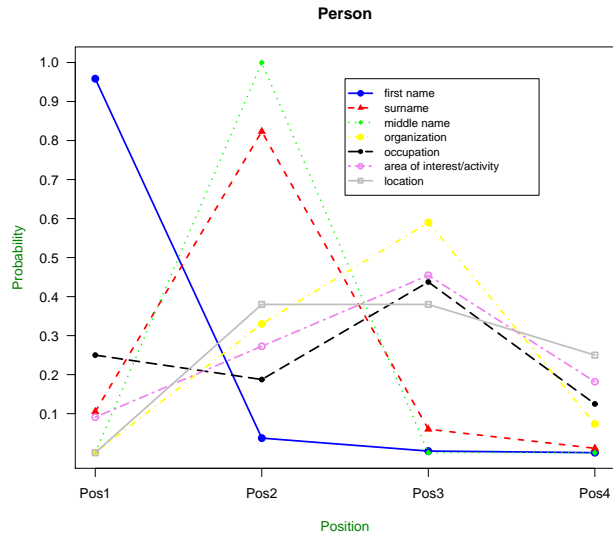


Figure 8.7: Probability distribution of attribute types for the first four positions in queries about Person.

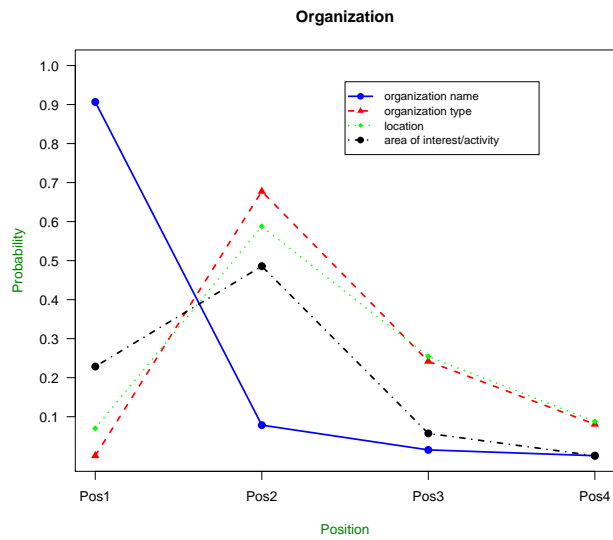


Figure 8.8: Probability distribution of attribute types for the first four positions in queries about Organization.

gesting that the two queries are indeed more likely about to different entity types. The analysis of the position distribution shows that queries about persons are more likely to present “first name” and “surname” in first and second positions respectively. If “organization name” is specified, this attribute type is more likely to occupy the third position. This order distribution match that of  $Q1$ . On the contrary, queries about organizations are more likely to present the name of the organization in first position, as we find in  $Q2$ . The position results suggest that  $Q1$  is more likely to be a query about a person, whereas  $Q2$  is more likely to refer to an organization.

Based on these results, we propose to extend the Bayesian model of attribute relevance presented in section 8.2.2 to incorporate position dependencies. We call this model Extended Bayesian Model (EBM). The Naive Bayes Model assumes the positional independence for attribute types: the conditional probability of an attribute type given an entity type is independent from the position of the attribute in the query. To incorporate position dependencies, we suggest to weight the probability  $p(E = e_k|A = a_s)$  by the position term  $Pos_j$  defined in Equation 8.13, as follows:

$$p(E = e_k|A = a_s, pos_j) = \frac{p(e_k) * p(a_s|e_k)}{\sum_{i=1}^m p(e_i) * p(a_s|e_i)} * Pos_j \quad (8.14)$$

where  $Pos_j$  is the probability of observing the attribute  $a_s$  in position  $j$ <sup>19</sup>. To compare the performance of the NBM with that of the EBM we tested the two models on the same sample of 125 queries randomly extracted from the queries collected in the entity search experiment. The results in terms of Precision, Recall and F-measure are reported in Table 8.15 and 8.16.

Measures	Person	Organization	Event	Artifact	Location
Precision	0.72	0.87	1	1	0.85
Recall	1	0.91	0.91	0.66	0.96
F-measure	0.84	0.89	0.95	0.80	0.90
<b>Overall Precision</b>	<b>Overall Recall</b>		<b>Overall F-measure</b>		
0.86	0.89		0.88		

Table 8.15: Test-set Evaluation of the NBM.

Measures	Person	Organization	Event	Artifact	Location
Precision	0.85	0.89	1	1	0.96
Recall	1	0.91	0.92	0.72	0.95
F-measure	0.92	0.90	0.95	0.84	0.96
<b>Overall Precision</b>	<b>Overall Recall</b>		<b>Overall F-measure</b>		
0.93	0.90		0.91		

Table 8.16: Test-set Evaluation of the EBM

The results show that the extended model may sensibly improve the disam-

<sup>19</sup>When an attribute type is not observed in a given position, the relevance of it is multiplied by a small constant term to avoid the nullification of the relevance value.

biguation process compared with the original model, supporting the hypothesis that the the order of terms within a query conveys semantic meaning that can be exploited for the disambiguation process.

### Entity type disambiguation in Web Search

In the Section 8.2.3 we have provided empirical evidence that the NBM (or its extended version, i.e. EBM) yields good disambiguation performance on our test sets. In this section we investigate the beneficial impact of the entity type disambiguation approach for an entity-centric search system. To perform this analysis we used an entity-id lookup system (available at <http://api.okkam.org/search/>) provided by the Okkam project and we compared the performance of the system in three different search conditions (respectively with correct disambiguation, with wrong disambiguation and without disambiguation).

To this purpose, we used a set of fifty queries randomly selected from our experimental dataset. In the first condition, i.e. *correct disambiguation*, we submitted the queries to the system specifying for each query the correct entity type<sup>20</sup>. In the second condition, i.e. *wrong disambiguation*, we submitted the queries using the same search functionality used in the previous condition, but specifying a wrong entity type randomly chosen between those provided by the system. In the last condition, i.e. *default condition*, the queries have been executed without filtering the results by entity type. The number of correct results in the first 20 returned results and the ranked position of the first correct match have been used to calculate a measure of the performance of the system.

In order to test the impact of the disambiguation on entity-centric search we tried to answer the following questions:

- What is the impact of using Entity Type Disambiguation versus not using it?
- What the impact of errors in Entity Type Disambiguation on the search results?

In order to be able to differentiate performance we used three different measures:

- The precision  $\mathbb{P}$  of results, measured as the number of entities related to the query to the full set of entities returned by the search engine. The search results returned are 20 at maximum. If  $|C|$  is the number of correct

---

<sup>20</sup>We used the search functionality of the system that allows to filter the results by entity type.

Performance Measure	Disambiguation		
	Default	Correct	Wrong
Precision $\mathbb{P}$	12.13%	<b>18.69%</b>	12.24%
Ranking effectiveness $\mathbb{R}$	13.00	<b>12.78</b>	28.98
Overall Performance $\mathbb{F}$	11.92%	<b>18.54%</b>	1.14%

Table 8.17: Performance Measures. Best result in **bold**.

or plausible answers and  $|A|$  the number of all results returned then the following formula calculates  $\mathbb{P}$ .

$$\mathbb{P} = \frac{|C|}{|A|}$$

- The ranking effectiveness  $\mathbb{R}$ , measured as the rank of the most appropriate entity for the query. If there is not plausible answer entity in the results we apply the dummy value of rank 30 as the ranking effectiveness. Thus, higher is worse in the ranking effectiveness.
- The overall performance for a given query. The overall performance is calculated by the formula in the following equation.

$$\mathbb{F} = \frac{31 - \mathbb{R}}{30} \times \mathbb{P}$$

This equation weights the precision by a function of the rank of the most plausible answer in the answer set. A perfect query answer will have only plausible entities returned, i.e.,  $\mathbb{P} = 1$  and the rank of the most plausible will be  $\mathbb{R} = 1$ . In this perfect case  $\mathbb{F} = 1$  and  $\mathbb{F}$  diminishes in every other case, reaching zero when no plausible result has been returned ( $\mathbb{P} = 0$ ).

In Tables 8.17 we report the performance of the system related to precision  $\mathbb{P}$ , ranking effectiveness  $\mathbb{R}$  and overall performance  $\mathbb{F}$ . We found that the correct disambiguation produced an improvement on all the measures of performance. To test if the difference in overall performance was significant between the conditions we conducted a comparison between pairwise conditions. In Tables 8.18 we report the results of the analysis which shows the beneficial impact of correct disambiguation, compared to default and wrong disambiguation and the cost of wrong disambiguation compared to default and correct disambiguation.

To conclude these results support our initial hypothesis that the effectiveness of entity type disambiguation can significantly improve the quality of the search results of an entity-centric system. On the other hand, the impact of wrong disambiguation can be very high. Thus, only if a highly effective disambiguation is used, disambiguation can be beneficial. Promisingly, the results

Relative Performance			
Comparison	Correct>Default	Default>Wrong	Correct>Wrong
p-value	✓ (0.01)	✓ (0.02)	✓ (0.001)

Table 8.18: Relative performance. In parentheses the p-value of the one-sided, paired t-test. ✓:Performance inequality statistically supported,

reported in 8.2.3 show that a highly effective disambiguation is possible and, more interestingly for the purposes of this work, this can be obtained making explicit the semantics underlying search queries.

## 8.2.4 Discussion

The entity search experiment gave us the opportunity to investigate one of the main issues of the present work, i.e. how a cognitive study on identification can contribute to inspire possible solutions to some of the most crucial problems about entity identification in systems which manage information about individual entities.

We focused, here, on a specific problem, which we named Entity Disambiguation Problem, which is the problem to identify the target entity of a query and assign to it the correct entity type.

To this purpose, we have proposed a probabilistic model for Entity Type Disambiguation that infers an entity type from the type of attributes a user specifies in a search query. We have also showed how the disambiguation performance can be improved including aspects related to how people organize the attributes within the query (i.e. order of attributes). And finally we have provided evidence of the impact that entity type disambiguation may have for an entity-centric search system.

However, our approach does not address the issue of how to perform automatically the assignment of attributes to their corresponding attribute types. However, the attribute type disambiguation is not a simple task. For some attributes the disambiguation process can be performed applying thesaurus-based disambiguation methods (an example is the use of a thesaurus of first names or location names), but for other attributes the disambiguation is more challenging. This is mainly due to the lack of contextual information in Web search queries - i.e. queries are typically composed by two or three terms - which makes the application of many Natural Language Processing Techniques, such as methods for Named Entity Recognition<sup>21</sup>, difficult to apply in this specific context.

For these reasons, in the last chapter of this work we propose a solution for

<sup>21</sup>Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

automatic disambiguation of attribute types and we present a Web application for Entity Type Disambiguation which uses this solution.





## Chapter 9

# Tracing the Identity of Individual Entities

Because individuals can change some of their properties while persisting as the same individuals, the singular cognition system needs a function of tracking a changing entity across time. This means that the system has to maintain the referential link between an entity in the world and its cognitive representation across time and change in order to make the identification process effective. In this part of our work we focus on the problem of how people judge the identity of entities over time and change (e.g. how people decide that an entity at one time  $t_0$  or context  $c$  is the same entity at another time,  $t_1$  or context,  $c'$ ).

This cognitive ability is crucial to daily life. We need to correctly identify our unique individual car, dog or spouse relevant to our own existence and successfully track those individuals across time and change.

There are at least two distinct representational systems underlying this fundamental aspect of human cognition.

The first is perceptual and has been largely studied in the context of visual perception and infant cognition, exploring the principles by which the visual system segments the visual input in discrete objects and bind individual views of objects into dynamic representations which persist across time, motion, featural change, and interruptions (see for example [187, 121, 211, 41, 230, 8]). When I'm watching a dog playing in a park, my perceptual system is able to preserve the unity of this entity although neither its retinal size or its shape remain constant. The perceptual system is also capable restoring continuity that has been temporarily broken in the stream of sensory inputs. The dog that reappears after running behind a tree will normally be treated as the same individual which was seen to disappear, provided that the disappearance was

short and that the parameters of motion remain more or less constant. In this case, it is not necessary to know the identity of the entity in place (i.e. to activate the singular representation of the entity and recognize it) to guarantee the experienced continuity of that entity. In this sense, the process is not dependent on high level conceptual representations, but it is ensured by temporary pre-conceptual representations (see 3.1 for a discussion on the models which addressed this issue).

The second system is fully conceptual and deals with conceptual information that comes into play when the object continuity can not be ensured by the correspondence process which attempts to match a low-level temporary representation (e.g. an object file in the Kahneman's model [121]) to a particular object perceived in the immediately preceding moments. Saying that the system is fully conceptual does not mean that perceptual information is not involved in the object identity tracking. It means, instead, that a high level conceptual representation of the object is activated. Perceptual information can be part of this representation and can be the most important information which mediates the identification process in some situations. When I recognize a friend of mine at a party, perceptual details can be the first elements of the conceptual representation of him which ensure the individual continuity. However there are situations in which perceptual information is insufficient (or completely absent) to trace the identity of an object. Consider the following situation. While reading a news item about a traffic accident, I start to suspect that the person involved in the accident is a classmate whom I lost touch with a long time ago. I have to decide about the identity of this individual using only the information reported in the article. I know that part of this information may reflect changes that the person has undergone, and I have to use this knowledge to decide if the description in the article is compatible with what I remember about that person.

Sometimes identity judgments also entail the ability to choose between alternative descriptions. If you are searching for a friend on Facebook and you find two or more alternative profiles registered under the same name, you have to decide which profile refers to the person you have in mind, in spite of information that might not match what you remember about that person.

In these cases, perceptual information can not help to trace the history of the entity involved and higher level information about identity must come into play.

The focus of this study is on the conceptual system which mediate identity judgments.

In 3.3 we have described two alternative approaches which have addressed the problem of object identity in terms of a conceptual system. The first ap-

proach, i.e. sortalism, claims that certain concepts (i.e. sortals) may determine rules for individuating and identifying their category members. The concept of table, for example, may consist in part of rules for differentiating individual tables in a mass of tables and other objects and identifying each table over time. A sortal specifies which properties of an individual category member can change over time (and in what way) and which properties are fixed. From this perspective, knowledge of categories dictates identification rules for their exemplars.

An alternative approach has been recently proposed by Rips et al. [192]. According to this approach the ultimate basis for identity is not rooted in high level representations of categories (i.e. general concepts), but depends on causal forces that determine the continuity of objects through time. Causal laws govern the life course of individual objects and people make use of causal information to identify objects across time and change.

In a series of evocative studies, the authors examined the role of causality in identity judgments and have proposed a new model of object identity named *Causal Continuer Theory*. We have already described the model in 3.3.2. Here, we summarize the most important aspects of the model for the purposes of our research and we discuss how we used the model as general framework to study the functioning dynamics of singular concepts in two experiments.

The model attempts to describe the cognitive processes people go through when they have to decide whether an individual object,  $x_0$ , existing at one time is identical to one of a set of candidate objects,  $x_1, x_2 \dots, x_n$ , existing at a later time.

The model derives from a philosophical theory, i.e. the Closest Continuer Theory, proposed by Robert Nozick [174] as a theory of personal identity. The Nozick's theory suggests that the identical object to the original  $x_0$  is the one that is, in some ways, the closest to it. In the Rips et al.'s model this closeness is explained in terms of causal dynamics and therefore the model has been referred to as *Causal Continuer Theory*. The intuition is that "the continuer of the original object must be a causal outgrowth of that original" (p. 7). Causal continuity captures the intuition that people think of causes as central to object persistence and suggests that what makes two entities identical with each other is not based on superficial similarity or sortal membership, but rather on a deep causal connectedness.

While the first element of the model deals with causality, the second element deals with closeness. As we noted above, the model assumes that in determining a continuer, people do not select something that is arbitrary far from the original. If there are two or more objects at a later time that are close enough to the original, the theory specifies that only the closest of these objects is identical to the original. However, if none of these potential continuers is significantly

closer to the original than the others, the model predicts that indecision can be generated due to the competition between the candidates. Another aspect related to the closeness is that in determining a continuer, people can not to select something that is arbitrary far from the original. If the candidates are causally too far from the original there may be no object that qualifies as identical to the original.

In the causal continuer framework, the authors proposed a two-step decision process on identity judgments. 1) The first step deals with considering as potential candidates only those objects that are close enough to the original; 2) the second step consists of selecting, within the range of candidates, the closest object as the one identical to the original.

But how can a cognitive agent decide which are the possible candidates of the original and then determine which is the closest to it? Our intuition is that these processes can be performed by means of singular concepts. In our framework, singular concepts are organized structures of semantic features or attributes which store our information about the individuals they are about. In chapter 8 we argued that these features are of different importance in concept representation and that the most important features in a singular concept are those which absolve better the identity function, i.e. those that are more useful to discriminate an individual from other similar individuals (attributes with high dominance and high distinctiveness).

Since individuals change across time, an important aspect of the identity function deals with the *mutability* of features. An individual object, such as a person or a car, can undergo a variety of changes in its properties, whereas other property changes are not compatible with identity. Total disassembly and reassembly may be possible for a watch but not for a person. This distinction between possible and impossible changes for an individual then determines, at least in part, the identity of that individual.

Within a causal framework, we argue that a change in a property is considered compatible with the identity of an individual, only if there is a causal explanation which can justify the change. Singular concepts would mediate this evaluation. Our hypothesis is that the identity judgment requires a matching process involving the original representation of an individual and one (or more) representations of possible candidates (continuers). Representations of possible candidates are selected considering the causal distance with the original representation. A match is found and the identity is assigned, if a causal path is traced from the initial representation to one of the alternative representations. The causal path is given by a causal explanation which justify the transition from one representation to another, i.e. a causal explanation for a change in one or more features. Since more than one path can be found, the contender with

the strongest causal path is selected as the one identical to the original. In other words people decide that a singular representation about a target individual  $x_0$  at one time  $t_0$  belongs to the same object as a representation of it at another time  $t_1$ , if there is a causal link which explains the transition of the representation at  $t_0$  into the representation at  $t_1$  and this link is the strongest compared to other links which connect the original representation with representations of other contenders.

The goal of the present study is to test how well the model explains people's identity judgments. To this purpose we performed two experiments in which we used descriptions of individual entities (e.g., people or organizations) and asked participants to make judgments about the identity of the entities across changes in the descriptions.

In a first experiment (experiment 1) we collected "mutability ratings", i.e. ratings of whether an individual could still be the same individual given a change in one of its features and "causality" ratings of how easy it is to imagine a cause for such a change. The goal of this experiment was to collect quantitative measures of causal distance which could be used to test the causal model in the second experiment. By collecting the judgments of causality and mutability from separate groups of participants we aimed to test whether these judgments were independent. This was an important requirement for the use of causality ratings in the second experiment, as we will discuss later.

As we have just noticed, in the causal model, identity judgments involve a double comparison process: the identical object must be causally close enough to be the original and must be closer than other close enough alternatives. Therefore, to test the model, we needed a situation in which at least two contenders were available and participants were asked to make a choice between potential continuers which differed by causal distance from the original. In our experiment we included a situation of this sort, asking participants to make identity judgments between alternative descriptions which were simplified versions of decisions that people have to make in many real-world situations (e.g. select the correct profile of a friend on Facebook). Our aim was to predict participants' judgments of whether an individual is the same as one, both, or neither of two alternatives, where the alternatives differ from the original in the change of a feature. For example, given a person who is 5 feet 10 inches and is a lawyer, which of the following individuals is the same as the original: a) one who is 6 feet 1 inch and is a lawyer, b) one who is 5 feet 10 inches and is an accountant, c) both, or d) neither.

Since our hypothesis assumes that the distance between two representations depends on how it easy to find a causal explanation which connects the two representations, the measures of causal distance collected in the experiment 1

were used to create the tasks of this second experiment. In this way we could use causal distance as predictor of the results in the experiment 2 and test the performance of the model.

The causal continuer model assumes that the identity decisions are context sensitive, i.e. an item that is closest in one situation may not be closest in another if the second situation contains an even closer object. Context sensitivity in identity has been largely criticized in philosophy [171] because it is seem implausible that the question of whether an object  $x_0$  is identical to  $x_1$  can depend on the presence of an individual  $x_2$  that may also exist at the same time as  $x_1$ . Of course, if we conceive identity as an intrinsic matter of an object, there is no room for contextual dependencies. However, we agree with the authors of the causal continuer theory that considering alternatives is an inevitable part of judging the identity of entities. This is especially true in information contexts where the identity decisions are usually performed in an information space rich of alternatives and people have to decide if a piece of information is about an entity target. Moreover the idea that an item in one situation may not be closest in another if the second situation contains an even closer object, is coherent with the identification process in an information system, since in this context the problem is to find the better candidate given a certain information context. Since our focus is on cognitive processes which are used in daily-life identity decisions with particular attention for situation which involve interactions with computer-based systems, we aimed to explore the cognitive plausibility of the contextual sensitivity assumption in identity judgments. To this purpose, the second goal of the study was to compare the performance of the causal continuer model with that of a simpler model, i.e. Naive Causal Model, which assumes that identity judgments are contextual-independent.

## 9.1 Experiment 1: Mutability and Causal Distance Norms Production

The aim of this experiment was to collect *mutability* ratings and *causality* ratings for features of individual entities and to test the correlation between the two measures.

The idea that features differ in their mutability has been first proposed by Love et al. [141, 225] to explain the centrality of a feature for a general concept. According to this view, the centrality of a feature represents the degree to which the feature is integral to the mental representation of a category, the degree to which it lends conceptual coherence. The authors have proposed an explanation of the feature centrality for general concepts based on the notion

of mutability. The idea is that the mutability of a feature in a concept is a measure of feature centrality, reflecting people’s willingness to transform the feature in a representation of an object while retaining the belief that the object is represented by the concept. Therefore the degree of centrality associated with a feature can be measured by asking people how easily they can transform their mental representation of an object by eliminating the feature or by replacing the feature with a different value, without changing other aspects of the object’s representation. For instance, when one thinks about robins, one envisions a creature that eats, builds nests, flies, has wings, a red breast, feathers, and so on. Nevertheless, one can successfully perform conceptual transformations in which one can imagine a robin that does not build nests but is still a robin. It could be more difficult to imagine a robin which lacked bones and still count as a robin. Features that are central to a representation, like “has bones” are referred as immutable, while those that are more easily transformed, like “builds nests” are referred to as mutable.

We argue that the notion of mutability can be used to quantify the relevance of a change in a feature for identity judgments. In this perspective, a feature is mutable of an entity to the extent that the feature can change without altering the object’s identity. The idea is that features of individual objects can be ordered according to their mutability using a task which requires to evaluate the probability that a change in one attribute occurs without changing the identity of the object.

The first goal of this experiment was to collect mutability ratings for features of individual entities using a measure which is a variation of one of the measures of mutability proposed by Sloman et al. [225] for general concepts, i.e. the easy-of-imaging measure.

In a typical easy-of-imaging task, subjects are asked how easily they could imagine an actual instance of a category without a specific feature. For example, how easily they could imagine “a real apple that does not grow on trees”.

Since we were focused on concepts of unique individuals, we adapted this task to the purposes of our study. First, we provided a brief description of an individual, presenting a profile composed by five attribute-value pairs. Then, we asked people to judge how easily could this individual still be the same if it were in all ways like that in the description except that one of the feature was changed in its value.

In this experiment we also collected a second type of ratings which we refer to as *causality* ratings. Since in our causal framework we assume that identity decisions are function of causal distance between singular representations, we were interested to quantify the degree of causal distance which divide two representations when these representations differ for the value of one attribute. In

other words, our aim was to quantify how easy is to imagine a causal explanation for a change of a representation on a specific attribute. To collect causality ratings we used the same method used for mutability ratings. We presented an attribute-value profile describing an individual and then we asked participants to judge how easy is to imagine a cause that can determine the change on a specific attribute of the description.

Mutability and causality ratings were collected from two different groups of participants, in order to ensure that the two measures (i.e. mutability and causality distance) were independent. This is because in the second experiment we aimed to use the causal distance as predictor of identity judgments and was important to exclude that causal judgments from which we derived the ratings involved a form of identity judgment.

On the contrary a type of identity judgment was involved in the mutability task. Since the mutability ratings asked how easily a feature could change while preserving identity, they already involve a type of identity judgment quite similar to that participants will be asked to make in the second experiment. Therefore, if the comparison between the two measures provided evidence for the independence of the two dimensions we could use the causality distance as a predictor for the identity judgments in the second experiment.

### **9.1.1 Method**

#### **Participants**

The participants were 32 Northwestern University students who took part in order to fulfill a course requirement in introductory psychology. Of the participants, 16 were randomly assigned to the mutability ratings group, the other 16 were assigned to the causality rating group.

#### **Stimuli**

In order to collect mutability and causality ratings we used descriptions of individuals by means of attribute-value profiles. Individuals were selected from the five categories (Person, Organization, Event, Artifact, Location) used in the previous studies on feature relevance. Five exemplars for each category were used resulting in 25 profiles. Each profile was composed by five attributes which were selected using the feature norms described in chapter 8. Since proper names are different from the other attributes of a singular concept - because of their nature of rigid devices of direct reference and unique markers of mental individuality - we decided to not include the names of the entities among the attributes tested for mutability and causality ratings. We collected ratings for each attribute of



the profile. Therefore, the experiment consisted of 125 trials resulting from five profiles for each category and five attributes for each profile. For example, a person profile used in the experiment was the following:

**NAME: Madison Smith**

AGE: 45

HOBBIES: tennis

OCCUPATION: reporter

PHONE: 202.287.3305

HEIGHT: 5'8"

The complete list of profiles used in the experiment is reported in Appendix D.1.

### **Procedure**

To collect the mutability ratings, participants were told they were presented with a series of descriptions of objects and that each object was described by a list of attributes. For each description they were asked to evaluate the probability that a change in one attribute occurs without changing the identity of the object. All the questions were of the following format: “How easily could X still be X if it had all the attributes of X except P?” where X is the specific object and it is changed in some manner with respect to the attribute P. To make these judgments, participants were asked to choose a number between 1 and 9, where 1 represented “very easy” and 9 represented “very difficult”.

For example, given the description above (which we refer as description A) a possible question was: “How easily could this individual still be the same if it were in all ways like that in the description A except that its OCCUPATION is changed? In other words, if you were presented with another description B that is in all ways like the description A except that the occupation is changed, how easily could these two descriptions refer to the same individual?”

Causality ratings were collected saying participants they were presented with a series of descriptions of objects composed by a list of attributes. For each description they were asked to imagine a possible cause that explains the change in one attribute. All the questions were of the following format: “How easy is to imagine a cause that changes the attribute A?”.

Mutability and causality ratings were obtained from two separate groups of subjects. Each subject provided ratings for all the trials of the corresponding condition. All trials were completely randomized between participants with the exception that the same profile could not appear in two consecutive trials.

At the end of the experiment, participants were asked to answer two questions about the strategies they used to provide their ratings.

The first question aimed to investigate whether participants took into account only the attribute type (e.g., “occupation”), without considering the specific value on this attribute (e.g., “reporter”) in order to provide their judgments, or whether they took into account both the attribute type and the specific value on this attribute. For example, to rate how easily the attribute “occupation” could change (or to imagine a cause that changes the occupation of the person) we wanted to know if they provided their judgments regardless of the fact that she was a reporter or they took in consideration the fact that she was a reporter.

The second question investigated whether participants took into account only the change in the specific attribute without considering possible interactions with other attributes, or whether they took into account both the change in the specific attribute and possible interactions with other attributes to make their judgments. For example, to rate the change in the attribute “occupation”, we were interested whether they judged how likely the person changes her occupation (or they imagined a possible cause that changes the occupation of the person) regardless of her “age”, or whether they considered also the age in making their judgments.

The motivation underlying these two questions was to verify the influence that the specific description we provided could have on the final ratings. Of course our goal was to minimize this influence and the instructions of the experiment were created to this purpose. In particular the main concern was about the interaction between attributes because we wanted to obtain context-independent ratings to be used in the second experiment.

### 9.1.2 Results

For each attribute the mean ratings across participants were calculated for each condition (i.e. mutability and causality). The complete list of attributes with the corresponding mutability and causality mean ratings is reported in D.2. Correlation across all the attributes were 0.66. The two measures of mutability and causal centrality correlates with each others. However, even though the correlation is quite high there is room for assuming that the two measure reflect, at least partially, independent constructs. For this reason we decided to use the causal distance measure as predictor for identity judgments in the second experiment.

The analysis of the questions about the strategies used by participants showed the efficacy of the instructions in reducing the influence of specific combinations of attributes and specific values on these attributes on the partici-

pants' judgments. The majority of participants reported that they took into account only the attribute type instead of the attribute-type combination to formulate their judgments, both in the mutability group (11 out of 16,  $p = 0.13$ ) and in the causality group (14 out of 16,  $p < 0.003$ ). In figure 9.1 we show the response frequencies in the two groups of participants.

We obtained a very similar pattern of results also for the second question about the response strategy used by participants. In particular, 12 out of 16 participants ( $p < 0.05$ ) in the mutability group and 13 out of 16 ( $p < 0.02$ ) participants in the causality group reported that they did not consider the interaction between different attribute types in the profile to formulate their judgments. Figure 9.2 shows the distribution of the response frequencies.

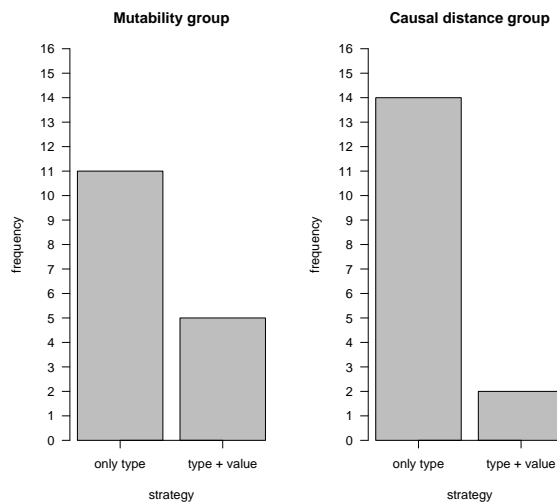


Figure 9.1: Response frequencies of the response strategies used by participants (question 1) in the two experimental groups.

## 9.2 Experiment 2: Identity Decisions across Change

The first goal of the second experiment was to test the predictions of the Causal Continuer Model using the causality ratings collected in the experiment 1 to estimate the causal distance of features in an identity decision task.

The second goal was to compare the performance of the causal continuer model with a simpler model, i.e Naive Causal Model, based on the assumption of contextual independence of identity decisions.

As we have noticed above, in the Causal Continuer Model, identity judgments involve a double comparison process: the identical object must be causally

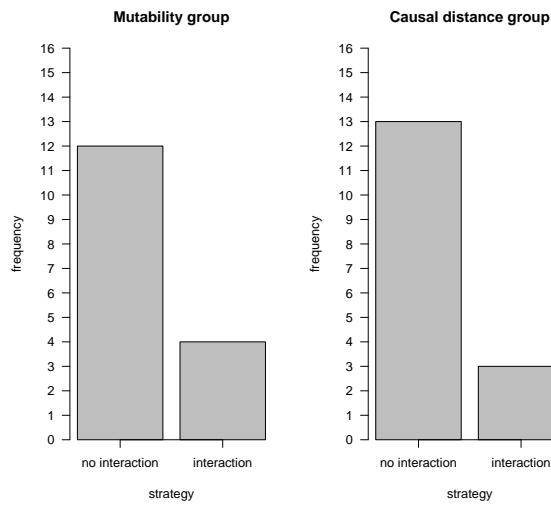


Figure 9.2: Response frequencies of the response strategies used by participants (question 2) in the two experimental groups.

close enough to be the original and must be closer than other close enough alternatives. To test how well the causal continuer model predicts people’s identity judgments, we needed a situation in which at least two contenders were available and the contenders differed for causal distance with the original.

To this purpose, we designed an experimental task that gave participants a choice between two alternative descriptions (i.e. continuers) which differed in terms of causal distance from an initial description of an entity (i.e. original). Then, we asked participants to judge whether the individual described in the original profile was the same as one, both, or neither of the two alternatives, where each of the alternative descriptions differed from the original in the change on a feature.

The goal of the study was to determine whether the causal continuer model could predict participants’ decisions about which description in each pair (or both or neither) was about the same individual as the initial one. We assume these decisions will reflect the model’s two-part structure: the participants’ notion of whether either alternative is causally close enough to be the original and also whether one alternative is causally closer than the other.

## 9.2.1 Method

### Subjects

45 undergraduate or postgraduate students of the Northwestern University of Chicago (USA) participated in the experiment in exchange for either course credit or a small payment.

### Procedure

To find out how well the causal continuer theory handles people’s identity judgments, we designed an experimental task that gave participants a choice between two potential continuers and varied the causal distance between the continuers and the original object. Participants were presented with the profile of an individual (i.e. a description containing two attribute-value pairs) and two alternative profiles (i.e. continuer 1 and continuer 2) that differed from the original by just the value of one attribute, but were identical to the original by the value of the other attribute. One continuer differed from the original on one attribute, the other continuer differed on the other attribute.

The participants’ task was to decide which of the alternative descriptions referred to the same individual as the original description. For each trial participants were asked to chose one of the following answers: “Only the continuer c1 refers to the same individual described in the original profile”, “Only the continuer c2 refers to the same individual”, “Both refer to the same individual” or “Neither c1 nor c2 refer to the same individual”.

For example, given the description of a person who lives in Germany and is 5 feet 6 inches, the question was to decide which of the following descriptions referred to the same individual as the original: a) one who lives in Ireland and is 5 feet 6 inches (continuer 1), b) one who lives in Germany and is 5 feet 3 inches (continuer 2), c) both, or d) neither.

Since we wanted to manipulate the causal distance between the continuers and the original description and study how well the model predicts the response distributions of participants, it was important to test different combinations of causal distance along the range of possible distances, varying from 1 (minimum distance) to 9 (maximum distance).

To this purpose, we divided the attributes of the categories used in experiment 1 in 5 quantiles on the basis of the mean causal distances. In this way, we created 5 sets of attributes for each category (i.e. Person, Organization, Event, Artifact and Location) ordered from the minimum to the maximum causal distance. Then, for each category we combined pairs of attributes which could be randomly extracted from the same set resulting in five pairs (i.e. the first pair from the first set, the second from the second set and so on), or from different

sets resulting in ten pairs, one for each combination of sets (i.e. the first with the second, the first with the third and so on). As a result we obtained 15 tasks for each category resulting in 75 tasks which were used in a within subject design. The procedure used to select the tasks ensured a good distribution in terms of causal distance which was reflected in an homogeneous distribution of the differences of the causal distances between the continuers. A graphical representation of this distribution for the Person category is reported in Figure 9.3.

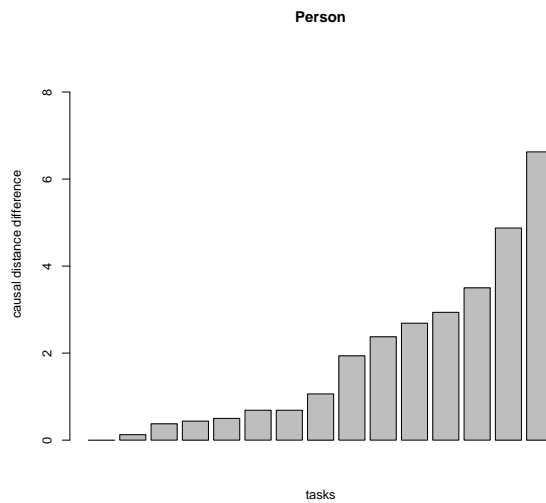


Figure 9.3: Distribution of the causal distance differences between the continuers in the person tasks used in experiment 2.

We note that we decided to use descriptions containing only two attributes in order to minimize possible interactions between attributes. For instance, a change on the attribute “occupation” can depend in some way from the attribute “age” (e.g. the fact that a 50 years old person is less likely to change her occupation compared to a 30 years old person). However, since the selection procedure selected randomly the combination of attributes from the quantile sets, possible dependencies between attributes within the same pair were evaluated separately from 4 judges who analyzed the tasks after they were automatically generated. In case a dependency was observed, the extraction procedure was repeated until obtaining an acceptable combination of attributes. The complete list of the profiles used in the experiment is reported in D.3.

## 9.2.2 Results

According to the Causal Continuer Theory, participants' responses on a particular trial should depend on a two-phase decision process. First, they need to determine whether either alternative description is close enough to the original to qualify it as referring to the same individual. Second, they need to determine whether one of the description is causally closer than the other.

Our assumption in this experiment is that the causal closeness between the original and a continuer is quantifiable by how easy is to find a causal explanation which may explain the transition from the original description to the continuer's description. The more easy is to explain the change, the less is the causal distance.

We assume that this process corresponds to an equivalent process in semantic memory which involves a comparison between singular concepts. The singular concept of the original is compared with the singular concepts of the continuers and the first step of the process establishes whether a causal explanation can be found to link the original representation with that of the continuers. Subsequently, it is established whether one of the singular concepts of the continuers is more strongly connected to the singular concept of the original.

If a causal link is found for one or both of the continuers and the second step of the identity process reveals that there is a strongest link between the original and one of the continuer, participants should respond that only the closer continuer is identical. On the contrary, if both representations of the continuers can be linked to the original, but it is not possible to decide which is the strongest connection, participants should respond that both continuers are identical to the original. In all the other cases they should answer that neither of the continuers can be considered identical to the original.

Since our assumption is that the causal closeness between the original and a continuer depends on how easy is to find a causal connection between the corresponding representations, we used the causal distance ratings collected in the experiment 1 to estimate 1) the likelihood that one or the other continuer (c1 or c2) is causally close enough to be identical to the original and 2) the likelihood that one of them is closer to the original. The quantitative model is expressed by the following equations.

$$P(\text{c1 or c2 close enough}) = 1 - (1 - P(\text{c1 close enough})) * (1 - P(\text{c2 close enough})) \quad (9.1)$$

In the equation 9.1 the probability that c1 or c2 are close enough to the original to be considered identical to it is calculated as the probability of two

disjunctive events, i.e. c1 is close enough or c2 is close enough. Disjunctive events are events which will be considered successful if at least one event is a success, therefore the probability that c1 or c2 are close enough is calculated as 1 - (the probability of both the two events NOT occurring). To estimate the probability P(c1 close enough) and P(c2 close enough) we used the causality ratings collected in the experiment 1 as follows:

$$P(\text{c1 close enough}) = \frac{(9 - \text{mean causal distance of c1})}{8}$$

$$P(\text{c2 close enough}) = \frac{(9 - \text{mean causal distance of c2})}{8}$$

We note that the causality ratings were provided on a 9-point rating scale (1=very easy; 9=very difficult). Therefore, the formula transforms the ratings into probability values ranging from 0 (when the causal distance is equal to 9) to 1 (when the causal distance is equal to 1).

The second step of the decision process is to establish whether the causal distance between the continuers is enough to consider one of them closer to the original. The assumption here is that only the continuer with the lower causal distance is the potential candidate to be the only one continuer closer to the original. If c2 is the continuer with the higher distance, the model predicts that c2 can never be closer than c1 and participants will never say that only c2 represents the original object.

Therefore if, for instance, c1 is the continuer with the lower causal distance, the probability that c1 is the closest continuer can be calculated as follows:

$$P(\text{c1 closer}) = \frac{P(\text{c1 close enough}) - P(\text{c2 close enough})}{1 - P(\text{c2 close enough})} \quad (9.2)$$

The equation 9.2 indicates that more the causal distances of the continuers are close, less likely c1 is considered the only closer continuer. Combining the equations 9.1 and 9.2 gives us the predictions for the identity judgments in experiment 2.

For example, assuming that c1 is the continuer with the lower causal distance the model predictions are:

$$P(\text{c1 identical}) = P(\text{c1 closer}) * P(\text{c1 or c2 close enough}) \quad (9.3)$$

where P(c1 identical) is the predicted probability that participants should identify only c1 as identical to the original.

$$P(\text{both identical}) = (1 - P(\text{c1 closer})) * (P(\text{c1 or c2 close enough})) \quad (9.4)$$



where  $P(\text{both identical})$  is the probability of a “both” response.

$$P(\text{neither identical}) = 1 - P(\text{c1 identical}) - P(\text{both identical}) \quad (9.5)$$

where  $P(\text{neither identical})$  is the probability of a “neither” response. As we have mentioned before, according to the model  $P(\text{c2 closer})$  is 0 in this case, because c2 can not be the only continuer chosen given that c1 is at a lower causal distance than c2.

To evaluate the model, we fit the model predictions to the percentage of responses of the experiment 2, using least squares approximation. Since we found a certain variability between the categories of entities (i.e. Person, Organization, Event, Artifact and Location) used in the experiment, we performed the analysis separately for the different categories. In Figure 9.4 and 9.5 we reported the percentage of responses obtained in the experiment 2 for the 15 trials of the person category. Lines with circle points represent the observed responses that the continuer c1, the continuer c2, both continuers or neither continuers refer to the same entity as the original description. Red lines with square points denote the model predictions (i.e. predicted percentage of responses). The graphs for the other categories are reported in Appendix D.4.1. Table 9.1 shows the overall fit of the model for each category.

Category	Model Fit ( $R^2$ )	Residual Standard Errors	gdl
Person	0.82	8.97	58
Organization	0.68	10.62	58
Event	0.32	13.19	58
Artifact	0.69	11.13	58
Location	0.77	10.19	58

Table 9.1: Causal Continuer Model Fit

The overall fit of the model is quite good for four out of five categories (i.e. Person, Organization, Artifact and Location), as we can observe from the  $R^2$  values<sup>1</sup>.

However, the model performs significantly worse for the category Event ( $R^2 = 0.32$ ). This result opens interesting questions about the ontological nature of events and the strategies used by people to trace the identity of events across time and change. We will discuss this aspect in section 9.2.3.

A second goal of the analysis was to compare the performance of the Causal Continuer Model with that of a simpler model which we refer to as Naive Causal

<sup>1</sup> $R^2$  is a statistic that quantifies the goodness of fit of a model. It is a measure ranging between 0 and 1, and has no units. Higher values indicate that the model fits the data better.  $R^2 = 1$  indicates that the model fits perfectly the observed data.

Model. This model assumes that participants make their decisions on the basis of their separate judgments of whether the continuer c1 or the continuer c2 refer to the same entity as the original description. This assumption differs from the causal continuer idea in that there is no explicit comparison for closeness between the continuers as expressed in the Equation 9.2. Under this assumption, the Naive Causal Model makes the identity decisions context insensitive, in the sense that the judgment on one continuer is not dependent on the presence of other continuers which can be more or less closer to the original. This means that, if we represent the probability that c1 is close enough to the original as  $P(\text{c1 close enough})$  and the probability that the c2 is close enough as  $P(\text{c2 close enough})$  as in the previous model, the Naive Causal Model computes the probability that c1 is identical to the original, c2 is identical, both are identical or neither are identical as follows. Assuming independence between decisions:

$$P(\text{c1 identical}) = P(\text{c1 close enough}) * P(1 - \text{c2 close enough}) \quad (9.6)$$

$$P(\text{c2 identical}) = P(\text{c2 close enough}) * P(1 - \text{c1 close enough}) \quad (9.7)$$

$$P(\text{both identical}) = P(\text{c1 close enough}) * P(\text{c2 close enough}) \quad (9.8)$$

$$P(\text{neither identical}) = 1 - P(\text{c1 identical}) - P(\text{c2 identical}) - P(\text{both identical}) \quad (9.9)$$

Estimating the component probabilities from the mean causality ratings, as we did earlier, allows us to fit the Naive Causal Model to the data. The goodness of the model fit for the five categories used in the experiment can be observed in Table 9.2. The Naive Causal Model performs less well than the Causal Continuer Model in four of the five categories (i.e. Person, Organization, Artifact and Location), as we can observe comparing the  $R^2$  values of the two models. However, the opposite pattern of results was found for the category Event, where the Naive Causal Model does considerably better than the Causal Continuer Model. This seems confirm that people use different strategies to evaluate the identity of events compared to other categories of entities.

In Appendix D.4.2 we reported a graphical representation of the Naive Causal Model fitting for all the categories of the experiment.

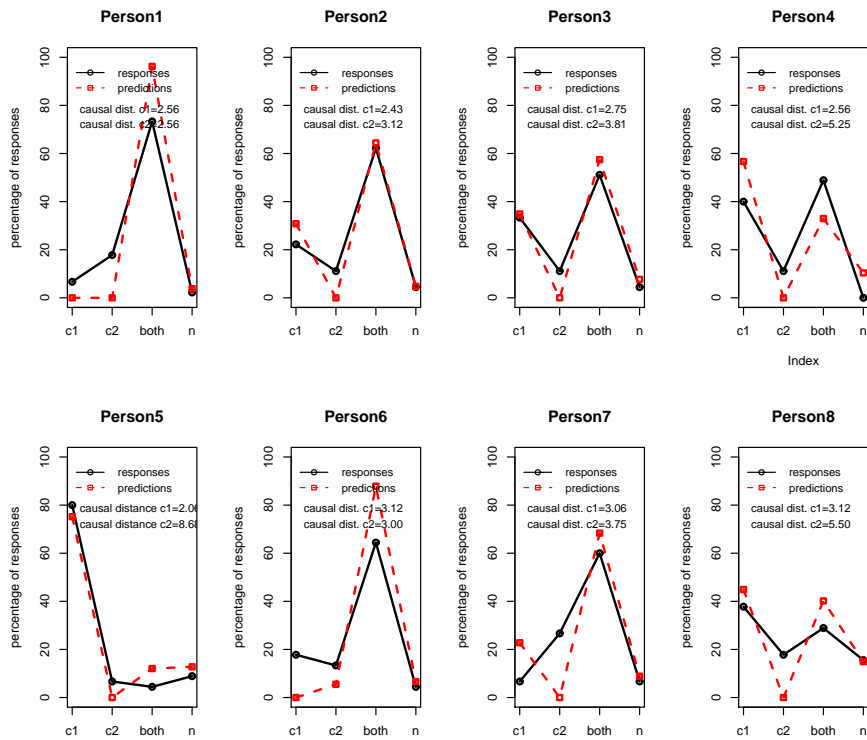


Figure 9.4: Person tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

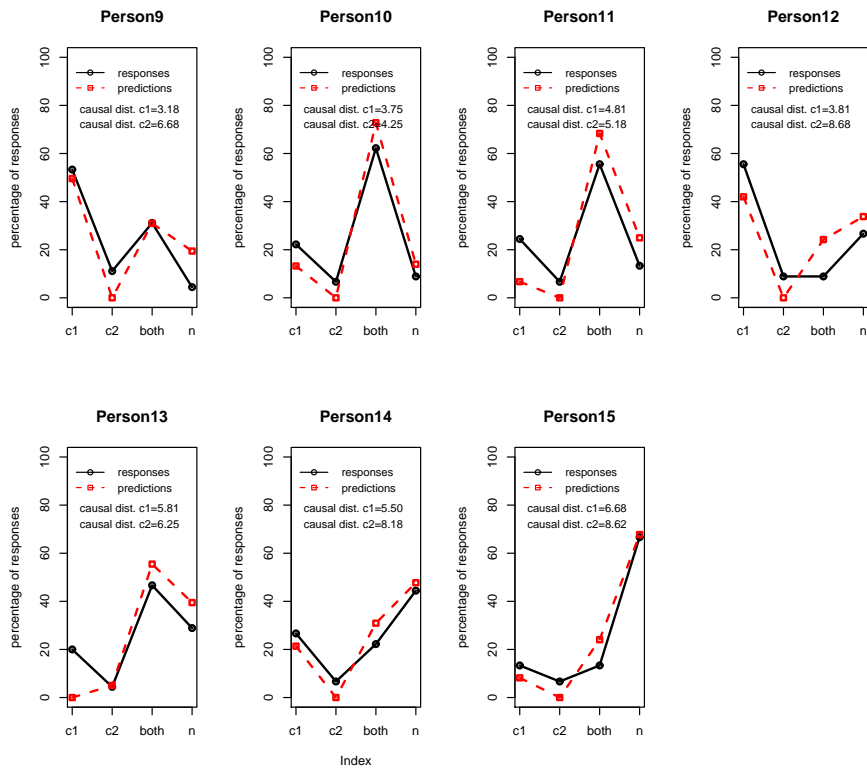


Figure 9.5: Person tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

Category	Model Fit ( $R^2$ )	Residual Standard Errors	gdl
Person	0.74	10.47	58
Organization	0.43	13.41	58
Event	0.65	9.50	58
Artifact	0.43	16.08	58
Location	0.38	15.85	58

Table 9.2: Naive Causal Model Fit

### 9.2.3 Discussion

In this experiment we have explored how people make identity decisions between alternatives across change.

Our goal was to test the hypothesis that people believe that causal forces come into play to determine the changes which objects can undergo. To decide if an individual  $x_0$  at one time and situation is the same individual  $x_1$  (changed on a certain aspect) at another time and situation, people would use causal explanations to evaluate the plausibility of the change. In other words, people use their knowledge about the probability that a certain cause may explain the change to evaluate the identity of objects across time or situations.

We know for example that a person can change hair style while still remaining the same person, but it is hard to believe that a dog can change its breed while remaining the same dog. This is because in the first case we can easily imagine a cause which explains that the person has changed hair style (e.g. she went to the hairdresser). On the contrary, it is difficult to find a causal reason which can explain that a dog changes its breed, at least in the real world.

The easy with which a causal explanation can be found for a certain change determines the causal distance of an individual  $x_1$  from the original individual  $x_0$ . Our hypothesis is that this causal distance is a general metric used by people to make identity judgments about objects of different categories. Since we believe that identity judgments are promoted by conceptual representations of individuals (i.e. singular concepts), we argue that causal distance is ultimately the metric for singular concepts. When a person has to decide, given a certain amount of knowledge about a target individual  $x_0$  at one time and situation, whether this individual continues to exist at another time and situation, she has to fix the referent of two singular concepts. More precisely, she has two singular representations in memory and she has to decide whether the two representations belong to the same object as a representation of it at two different times or situations. In order to perform such a process, the causal distance between the two representations is used. Since in many situations identity judgments also entail the ability to choose between alternatives, we argue that causal distance is also the metric used to discriminate between them. In particular we

hypothesize that the dynamics of functioning of singular concepts in promoting identity judgments between alternatives can be modeled by a Causal Continuer Model which assumes that these judgments involves a two-step process: the participants' notion of whether either alternative is causally close enough to be the original and also whether one alternative is causally closer than the other.

To test our hypothesis, we asked participants to make identity decisions between alternatives, using short descriptions of entities. The idea was reproduce simplified versions of decisions that people have to make in many real-world situations. In each task we presented a description of an individual entity (e.g., a person, an organization, an event and so on). This entity was described by a small set of attributes that represented all the information that the participant knew about the entity at the time of the decision. We conceived the description as a sort of explicit representation of the content of the singular concept of that entity. Two alternative descriptions were also presented together with the original, corresponding, in terms of mental representation, to other two singular concepts. Each of these descriptions differed from the original by the value of one attribute and this change corresponded to a certain causal distance, estimated using the causal ratings collected in the experiment 1. The task of the participants was to decide which (if either) of the two alternative descriptions was likely to refer to the same individual described in the original description. In terms of the functioning dynamics of singular concepts the task required to find a match (if any) between the singular concept of the original and those of the alternatives.

Using the mean causality ratings collected in the experiment 1 as predictors for the identity judgments in the experiment 2 we tested the performance of a quantitative version of the Causal Continuer Model. The results of the experiment showed that the model fit was quite good for the majority of categories used in the experiment, indicating that the model can predict the identity judgments between alternatives and, in our perspective, it can reproduce the dynamics of functioning of the underlying singular concepts. This results is also confirmed by the evaluation of an alternative model, which we called Naive Causal Model, which showed lower performance on the same categories. In particular, since the main difference between the two models is related to the contextual dependency of the identity decisions (assumed by the Causal Continuer Model but denied by the Naive Causal Model), the better performance of the Causal Continuer Model produces an evidence in favor of the idea that considering alternatives is an inevitable part of judging or inferring the identity of objects.

However, there is a considerable exception in this scene that is represented by the category Event. From our analysis the model is not able to reproduce

the identity decisions for entities belonging to this category. This is also the only category in which the Naive Causal Model performs better ( $R^2 = 0.65$ ) than the Causal Continuer Model ( $R^2 = 0.32$ ).

In chapter 8 we have discussed the ontological nature of the five categories of entities that we have used in this experiment, i.e. Person, Organization, Event, Artifact and Location. An important ontological distinction that can help to understand the difference between events and other categories of entities is between *endurants* (also called continuants) and *perdurants* (also called occurrents). Endurants are entities that are “in time”, they are “wholly” present (all their proper parts are present) at any time of their existence. On the contrary, perdurants are entities that “happen in time”, they extend in time by accumulating different “temporal parts”, so that, at any time  $t$  at which they exist, only their temporal parts at  $t$  are present. Events are perdurant entities, whereas all the other entities that we have considered in the experiment are endurants. Endurants and perdurants can be characterized by whether or not they can exhibit change in time. Endurants can “genuinely” change in time, in the sense that the very same endurant as a whole can change a property at different times; perdurants cannot change in this sense, since none of their parts keeps its identity in time. Suppose for example that a person has the property of “being a student” at a time  $t$  and the different property of “being a lawyer” at a time  $t_1$ . In both cases we refer to the whole object, without picking up any particular part of it. On the other hand, when we say that a perdurant like “the football game” has a property at  $t$  like “was boring” (at the beginning) and an another property at  $t_1$  like “was exciting” (say toward the end of the game) there are always two different parts exhibiting the two properties.

In this sense an event can not change in time as a whole. The different ontological nature of events could explain a different strategy used by participants in the experiment. Since events happen in time and are composed by temporal parts which are different across time, it is difficult to compare the continuers to establish if one is more close than the other to the original. This is because the changes can affect different temporal parts. This could explain the better performance of the Naive Causal Model in fitting the data, because this model does not assume the comparison between the continuers, but predict that participants make their decisions on the basis of their separate judgments of whether one or the other continuer is identical to the original.

Despite these differences, the main contribution of this study is to show that causal reasoning is of central importance to judgments of individual persistence. Moreover, from the best of our knowledge, this study is the first that has attempted to explicitly quantify the causal distance between alternatives and use this measure as the predictor of a causal model to infer the identity judgments.

This aspect is important to confer validity on the theoretical model. Indeed, one of the main criticisms that was made to the authors of the Causal Continuer Model was that to be able to make the claim that causal continuity is the factor that accounts for participants' identity judgments, they would have had to provide, minimally, some measure of causal distance [190]. On the contrary, according to this criticism, they offered post hoc descriptions, in each of their experiment, of how the observed results could have been due to causal reasoning. In one of their experiments, for example, Rips et al. [192] used stories about hypothetical transformations, similar to those adopted in some philosophical discussions of identity, describing a machine that could copy and transfer objects from place to place on a particle-by-particle basis. Participants read stories depicting a lion named Fred, whose copied particles were combined in some proportion with particles from another lion or a tiger to create a new creature. Participants were asked to decide whether the resulting creature was or was not Fred. The assumption was that causal closeness in this experiment depended on the percentage of the copy's particles that derived from the original. In the stories, the copying machine was the causal mechanism that produced closeness by copying particles and transmitting them. However, it was argued by [190], we cannot exclude that in this experiment people used other strategies to infer the identity of the object, like for example a similarity criterion (more particles from the original is equivalent to more similarity).

In our experiment we have manipulated measures of causal distance collected from a different group of subjects in a previous experiment, and we have shown that the participants' identity judgments can be predicted from a causal model which use these measures to infer the percentages of responses. This seems a direct way to provide evidence in favor of the hypothesis that causal continuity is the factor that accounts for participants' identity judgments.

Several experiments on object identity used fiction scenarios to explore the cognitive processes involved in identity judgments (see for example [137, 23, 192]). The use of this scenarios was criticized by some authors [190], arguing that in fictional contexts people are willing to accept kinds of transformations (no matter how extreme they are) which would not be acceptable in real world situations. This was considered another reason to cast a doubt on the effectiveness of transformation studies to explore how people make identity judgments in real-world situations.

Therefore, another contribution of our research is to have applied the Causal Continuer Model to a decisional context that reproduces simplified versions of decisions that people have to make in many real-world situations. There are many contexts in which identity judgments entail the ability to choose between alternative descriptions and this is particularly prominent in informational con-



text like the Web. For example, if you are searching for a friend on Facebook and you find two or more alternative profiles registered under the same name, you have to decide which profile refers to the person you have in mind, in spite of information that might not match what you remember about that person. When you use a search engine, like Google, you have to decide which link (or links) returned by the system refers to the entity you are looking for and you have to base this decision on a limited amount of information contained in the small fragment of the Web page, named snippet, which summarize its content. In all these situations the identity decisions are based on a limited amount of information and involve a decision between alternatives. We believe that our study provides an important contribution to explain how people perform identity judgments in situations like those described above and provides a plausible account within a theoretical framework that can be used in different contexts.

For example, understanding how people make identity decisions between alternatives can provide interesting insights for the development of systems which have to perform these decisions automatically or which involve the interaction with real users.

In 5.3 we have described an example of one of these systems (i.e. an Entity Name System) and we have noted a parallelism between the functioning dynamics of singular concepts and those involved in the maintenance of entity profiles in the system. We have observed that one of the main requirement for the system is the life-cycle management of the entity profiles across time and change. We envision the possibility to adopt the notion of causal distance for the development of algorithms which decide about the identity of entities through time. For example, one of the function which must be performed by these algorithms is entity merging. We have noted that since new profiles of entities are continuously added to the system and entity representation change incrementally as a consequence of the updating process, the system is supposed to revisit its identity decisions, i.e. it has to check if given the current status of information in the repository, entity matching would still support the same entity identity decisions. As a result of such a process it might be detected that two entity representations (with separate identifiers) actually refer to the same real world entity, requiring corrective actions which produce a unified representation from the two initially separated profiles. This process is named entity merging.

In order to decide whether two profiles refer to the same entity, a measure of causal distance can be adopted and a two-phase process like that suggested by the Causal Continuer Model can be performed. A first step would select possible candidates (i.e. those that are close enough to a given profile), while a second step would establish whether one of these candidates is sufficiently

closer compared to the contenders to be considered identical to the original and consequently merged with it. The causal distance could be quantified in an indirect way. Since in this specific case the causal distance measures how easily causal forces may change an aspect of an entity represented by the value of an attribute without altering its identity, the causal distance can be estimated by the degree of mutability of the same attribute in the profiles of all the other entities of the same type in the repository. The idea is that if it is unlikely that causal forces may change the feature (e.g. gender) represented in the system as the value of an attribute, we should expect that the degree of mutability of this attribute is low for the majority of entities of the same type. This means that the attribute changes rarely in their profiles. Estimating the causal distances from the histories of the entities in the repository, these measures can be directly applied in the merging decision process.

This is just an example of how understanding the cognitive processes involved in human identity decisions can be exploited in technological contexts, showing the potential for a profitable dialog between the two research fields.

## Chapter 10

# An Application for Entity Type Disambiguation in Queries using RDF Triples as Knowledge-Base

One of the objectives of this thesis is to provide evidence of how a cognitive study on the problem of individual identification can provide contributions for the development of technological applications.

In chapter 8 we individuated a possible ground where we found the opportunity to address this issue, i.e. the Entity Type Disambiguation Problem in Web search queries. The general question was how to determine from a set of keywords the entity a user is after, and the type of this entity, in order to limit the search to information about this precise entity. This issue is relevant for the information retrieval community since entity type disambiguation can be used to fix a number of failure cases in the relevance based search engines, but is also particularly meaningful for searching in Semantic Web content, where “aboutness” is a central aspect of information modeling. Finally the issue has a particular resonance for entity-centric approaches to information and knowledge management in distributed systems like the Web.

Knowing about what we want to know something can help us limit the search space significantly and improve the quality of search results.

As a first step of the research, we investigated in a user study which kinds of attributes humans actually consider relevant to identify different types of entities during the search process. The first contribution of the study was to identify

patterns of attributes that reproduce recurrent strategies in entity searching. For example, a query that is of the form “first name surname city” is indicative of a person search.

Based on these results, we investigated the assumption that an entity type can be inferred from the attributes a user specifies in a query and we proposed a Bayesian model for Entity Type Disambiguation that explores this assumption. We found that the performance of the model was very good and encouraging and we provided evidence for the beneficial impact of the Entity Type Disambiguation approach on the performance of an entity-centric search engine. However, the approach does not address the issue of how to perform automatically the assignment of attributes to their corresponding attribute types. Of course if we aim to implement the model in a real application, this is not a trivial task.

In this chapter we propose a possible solution to the automation problem and we show how an approach derived from a cognitive study can inspire technological solutions.

Since the previous analysis on the queries collected in the entity search experiment showed that the majority of attributes in a query contain named entities, we propose here a simplified approach to the problem of Entity Type Disambiguation. This approach is based on the assumption that the entity type of a query can be inferred by disambiguating the types of the named entities in the query.

To investigate our assumption we propose a new method that automatically extracts and classifies the named entities in a query and then infer the entity type of the target of the query (i.e. the entity the user is looking for).

As far as we know, there are very few studies that have addressed the problem of named entity disambiguation in queries. Named Entity Disambiguation in query addresses for queries the same problem which traditionally has been addressed by Named Entity Recognition (NER) in natural language texts. NER is the task of processing a text and identifying certain occurrences of words or expressions as belonging to particular categories of Named Entities (NE), such as the names of persons, organizations, locations, expressions of times, quantities and others. Several approaches have been proposed in literature (e.g. rule-based, supervised, unsupervised machine learning approaches) and a number of cues are utilized to identify named entities in textual documents. These may include local cues such as affixes, orthographic cues (e.g. capitalization), part-of-speech (POS) tags (i.e. linguistic categories of words) and phrasal chunks (i.e. simple syntactic structures) or external cues such as external lookup lists of familiar names (i.e. gazetteers) or training corpora (typically used for machine learning approaches).

However, direct application of exiting NER methods to queries would not

perform well. This is because queries are usually very short and, therefore, contextual information (i.e. words surrounding a word), which usually helps the disambiguation process in texts, is very limited. Moreover, very often terms in queries are not in standard form (e.g., all letters are in lower case), and thus many features are not sufficient for performing accurate disambiguation.

For these reasons, we propose a new approach for entity type disambiguation which mine semantic annotated data to provide knowledge for disambiguating queries. The approach exploits a large data set of semantic metadata (RDF triples) - extracted from (Semantic Web) documents - as a repository of entity-related semantic knowledge which is used to extract named entities in queries and classify them in possible types.

The general idea is that the disambiguation process can be tackled exploiting the subject-predicate-object structure of RDF statements (or triples) used to describe resources<sup>1</sup> in the Semantic Web. In the RDF data model, the subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. Our approach focuses only on a subclass of RDF metadata namely RDF statements whose objects are literals (plain text strings, such as, for example, “Rome”, “Barack Obama” or “June 1th 2009”). In these statements the predicate, establishing the relation between the subject and the object, makes explicit the type of information specified by the literal object. For example, in a statement describing that there is a person whose name is Barack Obama, the predicate would specify that “Barack Obama” is the “person name” of the entity identified by the subject. To restate in other words, “Barack Obama” can be seen as the value of the attribute “person name”. From that we can infer that “Barack Obama” is a named entity whose type is Person.

The idea is that many terms that compose an entity query match literals (or part of them) of RDF statements. Therefore the predicates of these statements can be indexed (using a large data set of metadata) and mapped into a set of possible candidate entity types for disambiguation. For instance, in the previous example, the predicate “person name” would map to the entity type Person. In this way, any keyword or combination of keywords inside of a query can be searched in the index to get the most likely entity type. Once the set of candidate entity types have been returned for all the entities in the query, the most likely entity type for the whole query is selected.

Even though we tested the effectiveness of the approach on a subset of types of named entities (person, organization and location), the method is general

---

<sup>1</sup>in the Semantic Web, the term *resource* encompasses every thing or entity that can be identified, named, addressed or handled, in any way whatsoever, in the Web at large, or in any networked information system.

enough to be extended to more types of named entities (e.g. event or artifact) and to other attributes which do not contain named entities but other kinds of information (e.g. the profession of a person, the type of activity of an organization, the topic of an event and so on).

Finally, since with the rise of the Semantic Web more and more RDF data becomes available on the Web, we believe that the approach can be improved progressively enlarging and updating the RDF triple store from which the entity types are extracted, according to the new information represented on the Web. If a new named entity is semantically tagged within a RDF triple, this information is potentially available for the disambiguation process. Therefore, the proposed approach is less static than other NER methods based on extensive gazetteers - including lists of names of people, organizations, locations, and other named entities - or on manually annotated training corpora used in data-driven statistical approaches.

## 10.1 Related Works

The performance of search engines depends on their ability to capture the meaning of a query most likely intended by the user. The intended meaning has been viewed so far as either a “topic” [105] or an “intent” [30]. In our approach we propose a third aspect of the intended meaning of a query based on the underlying “entity” that acts as the core of the information need (i.e. the target entity). This view is in line with a recent entity-centric vision of information and knowledge management on the Web.

The entity-centric aspect of the Web has been described and supported both conceptually [27] and in terms of implementation [57, 43]. We study the problem of Entity Type Disambiguation in the query terms, based on an already existing context of entity-centric search on Web data.

Users submit queries that usually contain a small number of keywords, i.e. the queries provide very limited information related to the intended information need. Query processing has been used as a way to infer the information need from the query. Several features additional to the query themselves have been used in the literature to describe a query related to the web, e.g., query log information, search context information [39] and anchor text [133]. Among the efforts of understanding the meaning of a query, we can distinguish between two main kinds of processing that have been applied on queries: *query segmentation* and *query classification*.

**Query Segmentation** aims to identify “meaningful” segments inside of a query, usually referring to word *collocations* (sets of words found to be neighbors more often than expected by chance, within a text corpus). This approach

essentially addresses a syntactic issue, because it does not identify the type of concepts and does not assign concept labels to segments. Two main approaches have been proposed in literature for query segmentation.

The first is based on mutual information (MI) between pairs of query words [194]. Risvick et al. [194] used data mining in query logs and document corpora to produce segment candidates and compute “connexity” measures. The general idea is to apply this database of segments and connexities to a query, for splitting it into segments according to a segmentation procedure. The procedure matches all possible subsequences of the given query in the segments database and assigns connexity value to each matched segment. Finally, it computes a segmentation score for each segmentation and return the most likely segmentation.

The second approach uses, supervised or unsupervised, machine learning techniques. Bergsma and Wang [18] proposed a data-driven, supervised approach to query segmentation. In this approach at each word position, a binary decision is made whether to create a segment boundary or not and the decision parameters are learned discriminatively from gold standard data. Tan and Peng [237] proposed an unsupervised method that uses a generative model (unigram model) to recover a query’s underlying concepts that compose its original segmented form. The model’s parameters are estimated automatically from a text corpus using an expectation-maximization algorithm.

The main difference between query segmentation and our approach for query disambiguation is that query segmentation separates a query into a number of units, but it does not identify named entities from units and also does not classify them into classes or types. However, we can consider the process of detecting named entities inside of a query as a specialized form of query segmentation and a prerequisite of entity disambiguation. As we will explain later, our approach for entity detection is similar to that of Risvick et al. [194] in that it uses an iterative segmentation procedure which matches all possible subsequences of a query to find the segment with the higher probability for a given type. In our approach, however, segmentation is not separate from classification but is an integral part of it.

A second kind of processing applied on queries is query classification.

**Query classification** is the task of assigning classes to whole queries, in order to improve the retrieval performance of search engines.

Query classification falls into two groups: 1) classification according to search intent, such as classification of queries into three general intent classes: informational, navigational or transactional intents [33, 201, 122]; 2) classification according to semantics of query, such as classification of user queries into a ranked list of predefined content categories [221, 15, 39]. Topical, i.e. topic-based, web search classification has been studied intensely, especially within the

KDD Cup 2005 competition <sup>2</sup>. The best performing system of the KDD Cup used an ensemble of classifiers using rich information from different sources (e.g., query, search engine related documents) to perform the classification [220].

Beitzel et al. [14], on the other hand, indicate that pre-retrieval (vs. post-retrieval) classification can be very effective, when the query category in the training set is assigned manually — and not determined through a bridging process that uses search engine-suggested categories as was done by [220].

In this work, we focus on entity-type classification and not topical classification. Furthermore, in query classification the whole query is classified and there is no further analysis on the internal structure of query. Instead, our aim is to reveal the internal semantic structure of query by classifying the entities inside of a query into types and inferring from these types the type of the whole query.

In this respect, our approach is more close to Named Entity Recognition (NER).

**Named Entity Recognition** is the task of identifying named entities in a written text and classifying them into appropriate entity types. Named entities are information units like names (such as person, organization and location names), temporal expressions (dates and times) and certain types of numeric expressions (monetary values and percentages). In the expression “Named Entity”, the word “Named” aims to restrict the task to only those entities for which one or many rigid designators, as defined by S. Kripke [128], stands for the referent. Since the most important rigid designators are proper names, early work formulates the NER problem as recognizing “proper names” in general. Overall, the most studied types are three specializations of proper names: names of persons, locations and organizations, collectively known as “enamel”. In this work we focus on these three main types of named entities.

Many approaches have been proposed for NER. While early systems were making use of handcrafted rule-based algorithms [76], modern systems most often resort to machine learning techniques, including supervised machine learning [22, 25, 5], semisupervised learning [32, 178] and unsupervised learning [51, 2].

Named Entity Recognition is usually performed on text documents and very few studies have addressed the problem of NER in queries. This is due the fact that queries are usually short and are often not well formed. Therefore, NLP techniques are difficult to apply in queries for high accuracy.

A first study that have recognized the importance of named entities in Web Search was conducted by Marius Paşca [177]. The author introduced a weakly supervised method for mining Web search queries in order to explicitly extract named entities, using templates. The main contribution of the study was to capitalize query data, instead of document collections, in order to explicitly

---

<sup>2</sup>Check <http://www.sigkdd.org/kdd2005/kddcup.html> for more information.



extract named entities that are expected to be relevant and suitable for later use (for example, to improve the quality of named entity recognizers to be used in Web documents). However, the intent was to extract named entities pertaining to various classes of interest to Web search users, rather than to classify individual entities inside of a specific query, as we envision in our study.

More related to our approach is a recent work by Guo et al. [97]. The authors proposed a probabilistic approach to the NER task in queries using query log data and a weakly supervised learning method referred to as WS-LDA (Weakly Supervised Latent Dirichlet Allocation). The approach exploits topic models (i.e. probabilistic models for uncovering the underlying semantic structure of a document collection) in a new application which considers contexts of a named entity as words of a document, and classes of the named entity as topics. The aim is to detect the named entities within query and find the most likely entity class given the context.

Our approach aims to address a very similar goal. However, there are important differences between the two approaches.

First of all the Guo’s method focuses on single named-entity queries (i.e. queries with contain only one named entity). A single named-entity query is represented as a triple including a named entity, a context and a class. The goal is to find the triple for a given query which has the largest joint probability. Our approach is more general in that it can handle more complicated queries with multiple named entities, inferring the type of the whole query from the combination of the all entity types within the query.

Secondly, the approach by Guo et al. employs weakly supervised learning using partially labeled seed entities and query log as external knowledge in an offline learning phase. The query log is used to provide patterns of entities, classes and contexts whose joint probabilities can be learned by the NER system.

On the contrary, our approach does not employ machine learning techniques. The proposed Entity Disambiguation method uses a data set of RDF triples as external knowledge that contain entity-related information. However this data set is not employed for the training process but is used to create an index that can be searched to find the most likely entity type given a certain keyword or combination of keywords. In this respect, the external knowledge derived from the RDF triple store is used as a lookup of terms and types rather than as a training corpora.

In a recent study Du et al. [59] proposed a method to overcome the lack of context information in queries. They proposed to utilize the search session information before a query as its context to address this limitation and improve two classical NER solutions which are known as Conditional Random Field

(CRF) based solution and Topic Model based solution, respectively<sup>3</sup>. The idea is to use the relationship between current focused query and previous queries in the same session to extract novel context aware features which are used to assign the most likely entity class to named entities. In the use of external knowledge and in the machine learning models adopted, this approach is more similar to that proposed by Guo et al.’s than that presented in the present work which aims to create a system that can recognize named-entities in a given query without prior training.

Named Entity Recognition is also at the core of a more specific line of research: personal name classification in Web queries. The task underlying personal name classification in queries is to decide whether a query is a personal name or not. Shen et al. [222] proposed an approach based on the construction of probabilistic name-term dictionaries and personal name grammars, which are used to predict the probability of a query to be a personal name.

An effort has been also made to identify and categorize queries that include geographical entities. Rocio Guillén [96] have proposed a method that combines information extraction (i.e. gazetteers) and patterns.

Compared to these approaches which focus on named entities of a specific type, the main contribution of our work is to propose a more general approach that can be potentially extended to disambiguate all the named entities in a query.

## 10.2 The Entity Type Disambiguation Problem

In this section we present again the formalization of the Entity Type Disambiguation Problem, as it was introduced in 8.2.2. We first describe the problem at a very general level. Then, we propose a simplification of the problem based on the idea of Named Entity Recognition in Query.

Without loss of generality, we can represent a query  $Q$  as a set of unknown terms  $T = (t_1, t_2, \dots, t_n)$ , each of which can be a single word or a combination of words. We assume that each term  $t$  specifies the value of an attribute  $a$ . For example, in the query “Barack Obama”, “Barack” is the value of the attribute “first name”, “Obama” is the value of “surname”. Assume that  $A = (a_1, a_2, \dots, a_n)$  is a set of predefined attribute types. We map every term  $t$  into one appropriate type in  $A$ . After this mapping is established,  $Q$  can be represented by a vector  $\mathbf{a}$  (an assignment of attribute types  $a_1, a_2, \dots, a_s$  to the

---

<sup>3</sup>A conditional random field (CRF) is a type of discriminative probabilistic model most often used for labeling and segmenting structured data, such as natural language texts or biological sequences.

A topic model is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents.

terms in  $T$ ). Finally, suppose that  $E = (e_1, e_2, \dots, e_m)$  is a small number of entity types.

The goal of Entity Type Disambiguation is to assign the most likely entity type  $e^*$  to a given query  $Q$  described by its attribute vector.

In 8.2.2 we have proposed a Bayesian Model for solving the Entity Type Disambiguation Problem. The model can be described by a classifier that is the function *disambiguate* ( $f : \mathbf{a} \rightarrow (E)$ ) that takes as argument a vector  $\mathbf{a}$  of attributes and returns the most likely entity type  $e^*$ . This function is defined as follows:

$$\text{disambiguate}(\mathbf{a}) = \arg \max_{e_k \in E} \frac{p(e_k) * p(\mathbf{a}|e_k)}{\sum_{i=1}^m p(e_i) * p(\mathbf{a}|e_i)} \quad (10.1)$$

The model assumes that if the attribute types are correctly assigned to the query terms, the target entity type can be inferred from the combination of attribute types. However, no methods have been proposed to automatically extract terms and assign attribute types to them. Moreover, the attribute types and their granularity have been decided a priori.

The main goal of the present study is to provide a possible solution to this problem. The proposed approach is general enough to address the Entity Type Disambiguation problem as formulated above. However, here we propose a simplified formalization of the Entity Type Disambiguation problem which considers only a subset of attribute types that is attribute types that contain named entities.

### 10.2.1 A simplified version of The Entity Type Disambiguation Problem

Restating the problem of Entity Type Disambiguation, the general idea is that the identification and classification of terms in a query can lead to the detection of the target entity type (i.e. the type of entity the user is looking for) by inferring the entity type from the coexistence of different term types in the query.

The analysis reported in 8.2 on real-world queries, provided by a large sample of participants in an experiment performing an entity search task, showed that about 90% of the queries about person, location and organization contained the name of the entity target, along with possible other information. This means, for instance, that if a user is looking for information about the President of USA, it is more likely that he formulates the query using the proper name “Barack Obama”, eventually specifying additional information (e.g. President of USA), than using a definite description such as “The actual President of USA”.

Therefore, since the final goal of Entity Type Disambiguation is to understand the type of entity the user is looking for (i.e. the target entity), the problem can be reformulated in terms of detecting the target entity within a query and assigning to it the corresponding entity type. If a query contains a single named entity the problem is reduced to detect the only named entity in the query and assign the most likely type to that entity. For instance, in the query “Barack Obama President” there is a single named entity “Barack Obama”. In this case the disambiguation problem consists in detecting “Barack Obama” as a named entity and assigning the correct entity type to it (i.e. Person). Therefore, for the majority of single-named-entity queries, the Entity Type Disambiguation problem can be reduced to a Named Entity Recognition problem (see for example [97])<sup>4</sup>.

However, many queries contain more than one entity and the Entity Type Disambiguation problem can not be entirely reduced to a problem of named entity recognition in this case. Once the named entities are detected and classified, the disambiguation needs to discriminate between the target entity and the context entity/ies and assign the whole query to the entity type of the target. Consider for example the query “Barack Obama USA”. In this case the query contains two named entities corresponding respectively to Person and Location types. From that, it comes out that even when the disambiguation process is reduced to the disambiguation of the only named entities, a further inferential step is necessary to detect the target entity among the named entities within the query. From this premises, we can reformulate the Entity Type Disambiguation as follows.

A named entity query  $Q$  can be represented as a set of unknown terms  $T = (t_1, t_2, \dots, t_n)$ , each of which can be a single word or a combination of words. Some of these terms corresponds to named entities, while others specify other kind of information.

The goal of Entity Type Disambiguation is to assign the most likely entity type  $e_t^*$  to all the named entities in  $Q$  and then infer the most likely entity type  $e_q^*$  of the whole query from the combination of the entity types in  $Q$ .

The Entity Type Disambiguation process consists in three phases:

1. Entity Detection
2. Entity Disambiguation
3. Query Disambiguation

---

<sup>4</sup>Of course this is a simplification of the original problem and in 8.2 we have already discussed potential failures of this approach. See for example 8.2.3 for a discussion on some remarkable examples. However, the cost of automation in this specific context forced to accept some degree of inaccuracy.

The first phase of the disambiguation process consists of detecting the terms in  $Q$  which refer to named entities. We name this process *Entity Detection*. The next phase, named *Entity Disambiguation*, consists of assigning the most likely entity type to each named entity in  $Q$ . Finally, the last phase consists of inferring the entity type of the whole query (i.e. the type of the entity the user is searching information about). We name this phase *Query Disambiguation* to differentiate it from the phase 2. Of course in case of single-named-entity queries, phase 3 coincides with phase 2. At the core of the Entity Type Disambiguation Problem as formulated above there is the task of detection and classification of named entities in query. In the next section we propose a new approach to address this task in an automatic way. The approach has been implemented in a prototype application for entity type disambiguation.

### 10.3 A new approach for Entity Type Disambiguation

The core of our approach is based on the idea that the disambiguation process can be tackled exploiting peculiar characteristics of the RDF metadata used to describe resources in the Semantic Web.

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web which is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions named RDF triples or RDF statements.

The meaning of an RDF statement is that some relationship - defined by the RDF predicate - exists between the RDF subject and the RDF object. This relationship can be visualized as a node and arc diagram (i.e. graph) whose nodes are the subject and the object, while the arc represents the relationship between them.

RDF is based on the idea of identifying resources using Web identifiers (called Uniform Resource Identifiers, or URIs). The subject of an RDF statement can be either a URI or a blank node, both of which denote resources. Resources indicated by blank nodes are called anonymous and are not directly identifiable by a URI. The predicate of a triple is a URI which also indicates a resource, representing the relationship between a subject and an object. Finally, the object of a triple may be a URI, a blank node or a literal (a plain text string, such as, for example, "Rome", "Barack Obama" or "June 1th 2010"). We note that a literal may be the object of an RDF statement, but not the subject or the predicate. Unlike a subject or object, a predicate must always be a Uniform

Resource Identifier. A RDF graph representing a triple with a literal object is shown in Figure 10.1.

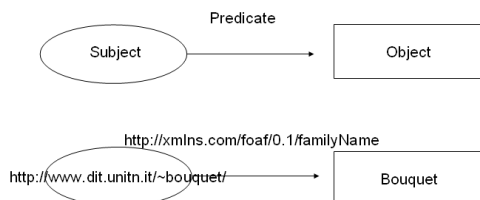


Figure 10.1: Graphical representation of an RDF statement.

A huge amount of RDF metadata are today available and a substantial part of these data are RDF statements whose objects are literals.

Our approach is based on the idea that many terms used in queries about specific entities could match literals (or part of them) of RDF statements. Since the predicate is the part of the statement that makes the object value a characteristic of the subject, the predicate conveys semantic meaning which identifies the type of information specified by the object. Of course, there's no way for a computer or a human to figure out what a specific predicate (i.e. URI) means, or how it should be used. This is where vocabularies and ontologies come in, describing explicitly the meaning and the relationships of predicates, as well as their domain of application. Consider, for example, the following RDF statement.

```
<foaf:Person rdf:about="http://disi.unitn.it/ bouquet/">
  <foaf:familyName>Bouquet</foaf:familyName>
</foaf:Person>
```

In the triple, `<foaf:Person rdf:about="http://disi.unitn.it/ bouquet/">` is the subject, `<foaf:familyName>` is the predicate, Bouquet is the object. The intuitive meaning of the statement is that there is a person (subject) whose surname (predicate) is Bouquet (object). However, the explicit meaning of the predicate (i.e. `foaf:familyName`) and its use is specified in the vocabulary of the corresponding ontology that is the FOAF ontology<sup>5</sup> in this case. The FOAF vocabulary specifies that “the `familyName` property is provided (alongside `givenName`) for use when describing parts of people’s names”<sup>6</sup>. From the definition of the predicate meaning and its entity type domain (i.e. Person), we

<sup>5</sup>FOAF is an ontology that has been designed to describe and integrate information about persons, their activities and their relations to other people and objects.

<sup>6</sup>Check [http://xmlns.com/foaf/spec/#term\\_familyName](http://xmlns.com/foaf/spec/#term_familyName) for more details.

can infer that a term labeled with the predicate “foaf:familyName” is referring to a named entity of type Person.

Therefore, using the specifications of RDF vocabularies, that make explicit the domain of application of predicates and their use, it is possible to map RDF predicates into a predefined set of attribute types or named entity classes.

From these premises derives the idea that a large data set of RDF metadata can be exploit to create an index of RDF predicates extracted from triples which contain literal objects. The index can be searched and used to extract possible candidate entity types (or attribute types) given a certain term of a query.

Since there is no restriction in the definition of predicates, different ontologies may use different predicates to specify the same relationship. For example, “http://www.w3.org/2006/vcard/ns#given-name” can be considered equivalent to “http://xmlns.com/foaf/0.1/givenname”. Moreover, also the same ontology may have more than one predicate which specifies the same relationship (e.g. foaf:givenName and foaf:firstName). Furthermore, for the purpose of our study, several predicates, even with a different semantic meaning, can be considered equivalent to disambiguate the general type of the object. A predicate that identifies the name of an author is considered equivalent to a predicate that identifies the name of a person, since authors are persons.

For these reasons, all the equivalent predicates returned by the index can be further mapped into a small number of attributes type or entity types to improve the efficacy of the disambiguation process.

The mapping of predicates into entity types used in our approach is reported in Appendix E.1.

We note that our approach is strongly dependent from at least three aspects: 1) the quality of the metadata used to create the index, 2) the availability of vocabularies that make explicit the use of predicates and 3) the discriminative power of the predicates, as specified in the RDF vocabularies. If the domain of application of an RDF predicate is too general, being used to specify a property that can be applied to more than a single type of entity, the predicate is not useful for disambiguation purposes. For this reason, we based the disambiguation process on a limited subset of the predicates extracted from the original data set that are the predicates which unambiguously refer to the entity types we are focused on, i.e. Person, Organization and Location. We remark that we started from these entity types mainly for practical reasons concerning the kinds, the amount and the quality of metadata today available on the Web along with the availability of ontologies and vocabularies underlying the use of named properties. However, we argue, the approach is general enough to be extended to other entity types in future, as the data set will be improved with new predicates and new mapping constraints.

## 10.4 PropLit: an application based on a index of RDF predicates

As we have introduced in the previous section, at the core of our approach lies an index of RDF predicates which is used to extract the candidate entity types.

To build the index, we used a data set composed of a billion of RDF triples<sup>7</sup> crawled during February-March 2009 based on datasets provided by Semantic Web search engines such as Falcon-S<sup>8</sup>, Sindice<sup>9</sup>, Swoogle<sup>10</sup>, SWSE<sup>11</sup>, and Watson<sup>12</sup>.

From the original data set we took in consideration only triples with a literal as object and we extracted only predicate-literal couples out of the data set. The result of this filtering operation produced 246 702 400 predicate-literal couples. The couples have been stored in an inverted index using Lucene<sup>13</sup>.

The index was build as a full-text *inverted index*. An inverted index is a mapping of words to their location in a set of documents. Most modern search engines utilize some form of an inverted index to process user-submitted queries. The goal of a search engine implementation is to optimize the speed of the query: find the most relevant documents where the keywords of the query occur. Once a *forward index* is developed, which stores lists of words per document, the index is inverted to create an *inverted index*. Querying the forward index would be highly consuming, in terms of memory, processing resources and time because it would require sequential iteration through each document and to each word to verify a matching document. Instead of listing the words per document in the forward index, the inverted index data structure is developed which lists the documents per word. In this way, the query can be resolved by directly accessing to the documents pointed by the corresponding words in the inverted index. Having determined which subset of documents or pages matches the query terms, a similarity (or ranking) score is computed between the query and each document/page based on the scoring algorithm used by the system. A largely used scoring algorithm is based on the *tf/idf* measure (term frequency-inverse document frequency) that evaluates how important a word is to a document in a collection or corpus, combining the number of times a given

---

<sup>7</sup>The RDF data set was provided for the Billion Triple Challenge 2009, an annual event for presenting new applications based on the Semantic Web vision. For more details see <http://challenge.semanticweb.org/>

<sup>8</sup><http://ws.nju.edu.cn/ontosearch/>

<sup>9</sup><http://sindice.com/>

<sup>10</sup><http://swoogle.umbc.edu/>

<sup>11</sup><http://swse.deri.org/>

<sup>12</sup><http://kmi-web05.open.ac.uk/WatsonWUI/>

<sup>13</sup>Apache Lucene is a free, open source information retrieval software library used for full-text indexing and searching in Java. See <http://lucene.apache.org/java/docs/index.html> for more details.



word appears in that document ( $tf$ ) with the inverse of the frequency of the word in the corpus ( $idf$ ). The relevance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word over the entire document corpus.

To build our index we used the same approach (an inverted index and a  $tf/idf$  measure of similarity), by mapping words to predicates, instead of words to documents. More precisely, our inverted index lists the predicates per word, including the position of the word within the literal. Namely, we treated the literals as texts of documents and the predicates as documents. Before creating the index, we used a filter which dropped out any stop words, words like articles, conjunctions, prepositions (a, an, the, and, of etc.) that occur so commonly in language that they might as well be noise for searching purposes. Just to make an example, given the literals  $L_0$ =Barack Obama,  $L_1$ =USA,  $L_2$ =Washington USA, we have the following full inverted index:

“Barack”:  $\{(0,0)\}$   
 “Obama”:  $\{(0,1)\}$   
 “USA”:  $\{(1,0), (2,1)\}$   
 “Washington” :  $\{(2,0)\}$

where, for instance, “Obama”:  $\{(0,1)\}$  means that “Obama” is in the literal  $L_0$  and it is the second word in the literal (position 1).

To measure the relevance of a given term for a specific predicate, we adapted the  $tf/idf$  measure to our context. We defined the *term frequency* of a term  $t_i$  for a predicate  $p_j$  as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where  $n_{i,j}$  is the number of occurrences of the considered term  $t_i$  (i.e. word or combination of words) in predicate  $p_j$  and the denominator is the sum of number of occurrences of all terms in predicate  $p_j$ . We note that we have many occurrences of the same predicate in our data set. Therefore,  $n_{i,j}$  is the number of times a given term  $t_i$  appears in all the occurrences of the predicate  $p_j$ .

We defined the *inverse predicate frequency* of a term  $t_i$ , as follows:

$$ipf_i = \log \frac{|P|}{|\{p : t_i \in p\}|}$$

where  $P$  is the total number of predicates in the data set and  $|\{p : t_i \in p\}|$  is the number of predicates where the term  $t_i$  appears.

From the combination of the two measures we obtain the  $tf/ipf$  measure as follows:

$$(tf - ipf)_{ij} = tf_{i,j} * ipf_i$$

A high weight in *tf-ipf* is reached by a high term frequency (for a given predicate) and a low predicate frequency of the term in the whole collection of predicates.

The index is the core module of an application for entity type disambiguation, which we named PropLit. PropLit provides two search functionalities: 1) Basic (without mapping) and 2) Advanced (with mapping).

The basic functionality of the index returns a list of ranked predicates for each search term. A search term can be composed by a single word (e.g. Barack) or a combination of words (e.g. San Salvador). When the search term contains a combination of words, the index returns the ranked predicates which contain the combination of words as if they were enclosed in quotation marks (i.e. in the exact order entered). A snapshot of the ranked list of predicates (first five results) for the terms “Barack” and “San Salvador”, respectively is shown in Figure 10.2 and in Figure 10.3

As we can note from the outputs reported in the figures, many predicates can be clustered since they convey the same semantic meaning (e.g. foaf:name and foaf:given-name) and a subset of them (those whose meaning is specific enough in the corresponding vocabulary) can be used to create a further mapping to a predefined set of entity types: person, location and organization<sup>14</sup>. Therefore, using the mapping reported in Appendix E.1, the index has been used to implement an advanced search functionality which maps the list of ranked predicates given a certain words (or combination of words) into a ranked list of entity types. The output of the advanced search for the term “Barack” is shown in Figure 10.4 and indicates that the term “Barack” is always object of predicates that map to the entity type Person. The numbers which accompany the entity type indicate the *tf/ipf* measure and its percentage value, compared to the other entity types.

Consider now the case of a query which contains two or more terms, like for example “Freddy Mercury” or “Paolo Bouquet Trento”. We have noted that in the basic search module described above, the index is searched using the combination of words typed in the search field, preserving the word order (e.g. “Freddy Mercury”). This approach presents, of course, a limit for the disambiguation process because it is dependent on the presence of the exact combination of words among the RDF objects of the data set. If I’m searching for “Carlo Bonatti” and no triples contain “Carlo Bonatti” as object, the

<sup>14</sup>The actual version of the index has been implemented to map into only three entity types because of the characteristics of the data set and the predicates available.

## PropLit INDEX

Query:

Query:

200 OK

```
{
  "Properties": [
    [
      "http://xmlns.com/foaf/0.1/name",
      [
        8140.7196999999996,
        "percentage: 37.63%"
      ]
    ],
    [
      "http://purl.org/dc/elements/1.1/title",
      [
        6306.4868139999999,
        "percentage: 29.15%"
      ]
    ],
    [
      "http://www.w3.org/2000/01/rdf-schema#label",
      [
        1190.21686,
        "percentage: 5.5%"
      ]
    ],
    [
      "http://www.w3.org/2006/vcard/ns#given-name",
      [
        1162.96,
        "percentage: 5.38%"
      ]
    ],
    [
      "http://www.w3.org/2006/vcard/ns#fn",
      [
        872.220000000000003,
        "percentage: 4.03%"
      ]
    ]
  ],
  ]
}
```

Figure 10.2: Snapshot of the list of ranked predicates returned by the basic search of the PropLit Index for the term “Barack”.

## PropLit INDEX

Query:

Query:

200 OK

```
{
  "Properties": [
    [
      "http://www.geonames.org/ontology#name",
      [
        1310.3311209999993,
        "percentage: 24.42%"
      ]
    ],
    [
      "http://www.w3.org/2000/01/rdf-schema#label",
      [
        1232.5302109999991,
        "percentage: 22.97%"
      ]
    ],
    [
      "http://www.geonames.org/ontology#alternateName",
      [
        1228.4354239999998,
        "percentage: 22.89%"
      ]
    ],
    [
      "http://purl.org/rss/1.0/title",
      [
        155.60181699999998,
        "percentage: 2.9%"
      ]
    ],
    [
      "http://xmlns.com/foaf/0.1/name",
      [
        147.41225800000001,
        "percentage: 2.75%"
      ]
    ]
  ],
}
```

Figure 10.3: Snapshot of the list of ranked predicates returned by the basic search of the PropLit Index for the term “San Salvador”.

## PropLit INDEX

Query:

Query:

200 OK

```
{
  "Barack": [
    [
      "person",
      [
        11011.777193000002,
        "percentage: 100.0%"
      ]
    ]
  ]
}
```

Figure 10.4: Snapshot of the output returned by the advanced search of the PropLit Index for the term “Barack”.

index does not return any result. Moreover, the approach can not be used to disambiguate more than a single search term. To overcome this problem, in the advanced module a different search algorithm has been used. The query processing in the advanced module has 4 main steps.

1. Tokenization
2. Stop word removal
3. Query representation
4. Query term weighting
5. Entity type mapping

1) As soon as a user inputs a query, the search module tokenizes the query stream, i.e., break it down into single terms (*tokenization*). A single term is an alpha-numeric string that occurs between white space and/or punctuation. In the query “Chicago USA”, for instance, the single terms are “Chicago” and “USA”.

2) The second step (*stop word removal*) removes all the stop words among the single terms obtained in the previous step<sup>15</sup>. If the tokenization process of the query “The Eiffel Tower” produces the following tokens “The”, “Eiffel” and “Tower”, the stop word removal eliminates “The” from the search terms.

3) The third step (*query representation*) creates a representation of the query containing single terms and sequences of terms which are used to search the Index. The goal of this step is to identify the meaningful units within the query<sup>16</sup>. These units not always coincide with single terms. Consider for example the query “New York USA”. The query contains three single terms “New”, “York” and “USA”, but in this case there are only two meaningful units “New York” and “USA”. To identify the meaningful units within a query, each single term is individually submitted to the Index, as well as each possible sequence composed by a single term and the terms that follow it within the query. For example, if we have a query  $Q$  of 3 single terms and we represent this query as a vector  $Q = (t_1, t_2, t_3)$ , we generate the following sequences:  $s_1 = (t_1, t_2)$ ,  $s_2 = (t_1, t_2, t_3)$ ,  $s_3 = (t_2, t_3)$ . As we can understand from the example, these sequences do not correspond to all the possible combinations of words which compose the query (i.e. the sequence  $s = (t_1, t_3)$  is not included) because we adopted the restriction that a sequence must be composed by contiguous words.

---

<sup>15</sup>A stop word list typically consists of those word classes known to convey little substantive meaning, such as articles (a, the), conjunctions (and, but), interjections (oh, but), prepositions (in, over), pronouns (he, it), and forms of the “to be” verb (is, are). Stop words are removed based on this list.

<sup>16</sup>A meaningful unit in our approach is a unit that contains a named entity.

In the previous example, the terms and sequences of terms submitted to the index are the following: “New”, “York”, “USA”, “New York”, “York USA” and “New York USA”. The assumption is that some terms (e.g. USA) or term sequences (e.g. “New York”) should be more represented among the RDF objects of the data set compared to others (e.g. “York USA”) and receive higher ranking scores in the next step. However, when a given meaningful sequence of terms, such as the first name and surname of a person (e.g. “Carlo Bonatti”) is not present among the objects of the data set (i.e. there are no RDF triples about that specific person), the disambiguation can be performed anyway combining the disambiguation of single terms. In the example, it is likely that there are many triples which are about persons named “Carlo” and in the same way it could be that there are triples about persons with the surname “Bonatti”. Hence, even though there are no triples about the specific person which we are looking for, we can disambiguate the two single terms as referring to a person. The example is shown in Figure 10.5.

4) In the fourth step (*query term weighting*), terms and sequences of terms are submitted to the index using the basic search module and the corresponding lists of ranked predicates, according to the *tf/ipf* measures, are obtained.

5) Finally, using a mapping function which maps predicates to entity types according to the mapping schema reported in Appendix E.1, a ranked list of entity types for each term and sequence of terms is returned (*entity type mapping*). Having assigned the ranking scores, the terms and/or the term sequences with the highest scores provide the term disambiguation returned by the Index. In Figure 10.6 is shown the output of the advanced search module for the query “New York USA”. We circled in red the suggested disambiguation according to the highest ranking scores. We can note that the system correctly identifies the meaningful units “USA” and “New York” and assigns the correct entity type to these units.

## 10.5 Index Evaluation

When we have introduced the Entity Type Disambiguation Problem, we noted that the problem can be reduced to an entity recognition problem in case of queries which contain a single named entity. Therefore, as a first evaluation of the application we conducted an analysis on a sample of queries containing a single named entity, such as “Barack Obama”, “New York” or “IBM”. The queries were randomly extracted from those collected in the entity search experiment described in 8.2 and the evaluation was performed manually. We tested the index on sixty queries for each entity types (i.e. person, organization and location). Each query was submitted to the advanced search module. For eval-

## PropLit INDEX

```
Query:  
Query:  
200 OK
{
  "Properties": [],
  "Number of hits": 0
}
```

(a) basic

## PropLit INDEX

```
Query:  
Query:  
200 OK
{
  "bonatti": [
    [
      "person",
      [
        1320.2643585,
        "percentage: 99.62%"
      ]
    ],
    [
      "location name",
      [
        7.3931255,
        "percentage: 0.38%"
      ]
    ]
  ],
  "carlo": [
    [
      "person",
      [
        23110.093104999982,
        "percentage: 99.14%"
      ]
    ],
    [
      "location name",
      [
        264.55121999999998,
        "percentage: 0.9%"
      ]
    ]
  ]
}
```

(b) advanced

Figure 10.5: Output of the basic (a) and advanced (b) modules for the query “Carlo Bonatti”

uation purposes, when multiple results were returned by the index for a given query, we considered the answer with the highest  $tf/ipf$  value. However, when a disambiguation was returned for a sequence of terms, we took in consideration the corresponding outcome for the evaluation, instead of considering the single terms (see Table 10.1 for an example), even though single terms presented higher  $tf/ipf$ . This is because we assume that a match with a sequence of terms is more relevant for disambiguation purposes than a match with single terms.

## PropLit INDEX

Query:

Query:

200 OK

```
{
  "USA": [
    {
      "location name": [
        [
          34475.175130000003,
          "percentage: 97.77%"
        ]
      ]
    },
    {
      "person": [
        [
          788.00405999999999,
          "percentage: 2.23%"
        ]
      ]
    }
  ],
  "York USA": [
    {
      "location name": [
        [
          1861.4016185,
          "percentage: 100.0%"
        ]
      ]
    }
  ],
  "New York USA": [
    {
      "location name": [
        [
          2524.52859300000002,
          "percentage: 100.0%"
        ]
      ]
    }
  ]
},
  "New York": [
    {
      "location name": [
        [
          65082.754280000000,
          "percentage: 91.98%"
        ]
      ]
    },
    {
      "person": [
        [
          5675.28935999999999,
          "percentage: 8.02%"
        ]
      ]
    }
  ],
  "York": [
    {
      "location name": [
        [
          39794.0272799999995,
          "percentage: 90.2%"
        ]
      ]
    },
    {
      "person": [
        [
          4322.00033000000002,
          "percentage: 9.8%"
        ]
      ]
    }
  ],
  "New": [
    {
      "location name": [
        [
          29821.1423000000003,
          "percentage: 90.14%"
        ]
      ]
    },
    {
      "person": [
        [
          3161.66740000000002,
          "percentage: 9.86%"
        ]
      ]
    }
  ]
},
}
```

Figure 10.6: Output of the advanced search module for the query “New York USA”

Consider, for example, a query like “George Washington”. Since “Washington”



can be the surname of a person or the name of a city, we expect that the disambiguation of the sequence “George Washington” is more relevant than the disambiguation of the single terms separately, even though the absolute value of  $tf/ipf$  may be lower compared to that of single terms<sup>17</sup>.

Term	tf/ipf		
	Person	Location	Organization
Hillary	7710.14	139.32	0
Clinton	4391.77	3862.97	0
Hillary Clinton	<b>345.19</b>	0	0

Table 10.1: Entity Type Disambiguation example. In bold the answer which we considered for the evaluation.

In Figure 10.7 we report the correct disambiguation frequencies (i.e. true positives) for each entity type, while the results in terms of *Precision*, *Recall* and *F-measure*<sup>18</sup> are reported in Table 10.2.

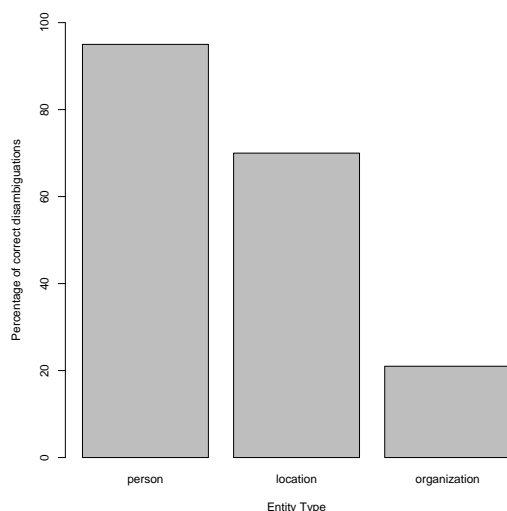


Figure 10.7: Percentage of correct disambiguations (true positives) on single term queries.

The results show that the system performs quite well in disambiguating

<sup>17</sup>This is because the  $n_{i,j}$  of a single term in the  $tf_{i,j}$  formula is  $\geq$  than  $n_{i,j}$  of a sequence which contains that term.

<sup>18</sup>*Precision* is the ratio of the number of queries correctly assigned to the entity type (true positive) to the total number of queries correctly (true positive) and incorrectly (false positive) assigned to that type. *Recall* is the ratio of the number of queries correctly assigned to the entity type (true positive) to the total number of queries that should have been assigned to that type (true positive+false negatives). *F-measure* is the harmonic mean of precision and recall.

Measures	Person	Organization	Location
Precision	0.55	1	0.81
Recall	0.95	0.22	0.70
F-measure	0.70	0.36	0.75

Table 10.2: Performance of the advanced search module on single term queries.

entities of Person and Location types, while it performs more poorly in disambiguating queries about organizations. We argue that one of the main reasons which can explain this significant difference concerns the characteristics of the data set we used to create the index. In Appendix E.2 we reported the list of the first 50 RDF predicates and the corresponding frequencies calculated on the original data set of RDF triples used to create the index, before filtering the data set (i.e. before extracting the triples with literal objects). This frequency distribution shows the most represented predicates in the data set and the ontologies to which they refer. We can note that the majority of these predicates refer to ontologies specialized to code information about persons (e.g. FOAF ontology) or about locations (e.g. Geonames). This may explain the significant better performance of our application in disambiguating queries about persons and locations than queries about organizations. The lower precision for the entity type Person is due to the fact that many terms which refer to organizations are mapped into Person type (false positives), because often rdf triples about persons contain the affiliation of the person along with the proper name. This aspect represents an element of noise in the data set due to an improper use of the semantics of the predicates. We believe that this element of noise should be overcome increasing the amount of rdf triples in the dataset uniquely referring to organizations. The imbalance in the predicate distribution in favor of Person and Location types, also explains the high precision for the Organization type which contrasts with the low recall.

Given the poor performance of the system in disambiguating named entities referring to organizations, we decided to limit the second phase of the evaluation to queries about Person and Location.

In the second phase of the evaluation we tested the system using queries containing multiple named entities and we evaluated two aspects of the performance of the application:

1. Entity Detection: the detection of the named entities within the query.
2. Entity Disambiguation: the assignment of the correct entity type to each named entity of the query.

In order to be able to differentiate performance on these two aspects we used the following two measures:

1. The *Detection Effectiveness*  $\mathbb{Z}$ , measured as the sum of the number of entities correctly detected in each query of the test set, divided by the number of entities that should have been detected/disambiguated in the query. The Detection Effectiveness has been calculated using the following formula:

$$\mathbb{Z} = \sum_i \frac{Ndt_i}{N_i}$$

where  $Ndt_i$  is the number of entities correctly detected in the query  $Q_i$  and  $N_i$  is the number of entities that should have been detected in the query  $Q_i$ .

2. The *Disambiguation Effectiveness*  $\mathbb{L}$ , measured as the sum of the number of entities correctly disambiguated in each query of the test set, divided by the number of entities that should have been detected/disambiguated in the query. The Disambiguation Effectiveness has been calculated using the following formula:

$$\mathbb{L} = \sum_i \frac{Nds_i}{N_i}$$

where  $Nds_i$  is the number of entities correctly disambiguated in the query  $Q_i$  and  $N_i$  is the number of entities that should have been disambiguated in the query  $Q_i$ .

Both measures range from 0 to 1, with 1 representing the maximum effectiveness. The analysis was performed on a test set of 128 queries (64 for each entity type) extracted from those collected in the entity search experiment. In Tables 10.3 we report the performance of the system related to Detection Effectiveness  $\mathbb{Z}$  and Disambiguation Effectiveness  $\mathbb{L}$ .

Performance Measure	Person	Location
Detection Effectiveness $\mathbb{Z}$	0.83	0.63
Disambiguation Effectiveness $\mathbb{L}$	0.68	0.66
N	64	64

Table 10.3: Detection and Disambiguation Effectiveness of PropLit on queries with multiple named entities. N = number of queries.

The Disambiguation Effectiveness for queries about persons is not significantly different from that of queries about locations ( $p = 0.79$ ), but the Detection Effectiveness is significantly ( $p < 0.01$ ) higher for queries about persons than for queries about locations. From our analysis it comes out that one of

the reasons that can explain this difference is due to the way in which semantic information about locations is coded in RDF triples. We found that very often two or more locations are included in the same predicate (e.g. Trento, Italy). This aspect may have reduced the effectiveness of the system in detecting the named entities within the queries favoring the detection of co-occurrences of entities.

In Section 10.2.1 we have proposed a simplified version of the Entity Disambiguation Problem and we have identified three main phases of it: 1) Entity Detection, 2) Entity Disambiguation and 3) Query Disambiguation. We noted that for single-named-entity queries, phase 2 and phase 3 coincide, that is disambiguating the type of the unique named entity in the query coincides with disambiguating the type of the whole query. Instead, for queries with multiple named entities when the entities within the query have been disambiguated, the type of the target entity must be inferred from the combination of the types of entities identified in phase 2. This phase is not directly implemented in the actual version of our system and represents an open field of future research. We believe that the investigation described in 8.2 may suggest useful insights into the implementation of this module as well as into the improvement of the entity disambiguation itself (phase 2). In order to define a baseline from which the impact of future solutions can be measured, we performed a final evaluation of the performance of the system, using a simple heuristic decision procedure that is to assign to a given query the entity type of the first entity of the query. For example, if we have a query  $Q$  like “John Lennon Beatles” and the entity disambiguation returns “John Lennon  $\rightarrow$  Person” and “Beatles  $\rightarrow$  Organization”, the heuristic procedure assigns the entity type Person to the whole query. In Table 10.4 we report the baseline performance of the system on the 128 queries used for the previous analysis.

Performance Measure	Person	Location
Precision	0.75	0.78
Recall	0.92	0.77
F-measure	0.76	0.84

Table 10.4: Baseline performance of Proplit on multiple-named-entity queries

The analysis shows that the heuristic procedure provides quite good results, indicating that for a substantial amount of queries the first entity is indeed the entity target. However, we have already discussed that this “rule of thumb” approach has the advantage of reducing the complexity of the disambiguation process, but has an important side effect. The success of the approach depends exclusively on the ability of disambiguating a unique piece of information within the query. If the process fails, the system will return a wrong disambiguation.

On the contrary, an approach such as that we have proposed in 8.2 which is based on the disambiguation of attribute types, instead of entity types, is less vulnerable to mistakes in disambiguating single attributes, because a certain pattern of attributes may still suggest the correct disambiguation despite the presence of a mistake. This is particularly true for the extended approach (i.e. extended version of the Bayesian Model) that takes into consideration the order of the attribute types within the query. We can illustrate the problem with an example. Consider the query  $Q$ ="Paris Hilton Hotel". If we submit the query to the advanced search module of our system we find that the best disambiguation returned by the system is "Paris Hilton  $\rightarrow$  Person ( $tf/ipf=346.50$ )". For a human is quite simple to understand that in this case the correct entity detection should be "Paris" and "Hilton Hotel" and that the type of the entity target should be Location rather than Person. This interpretation is suggested by the presence of the word "Hotel" that suggests that Hilton is the name of an hotel rather than the surname of a person. From this interpretation follows that Paris is more likely to be the name of a city than the name of a person. The example shows that the approach of disambiguating the first entity of the query may be not effective in cases like this and the main reason of the failure is that the processes of entity detection and disambiguation do not consider the dependencies between the sequences of terms.

We believe that the results of the entity search experiment can be used to overcome this problem. One of the main results of the investigation about the entity search experiment described in 8.2 showed that the disambiguation performance of the Naive Bayes Model can be improved extending the model to incorporate the position of the attribute types within the query. We argue that the same insight can be extended to the proposed simplified version of the Entity Disambiguation Problem. Going back to the previous example, the results of our experiment show that a pattern of attributes like "location name (Paris), location name (Hilton Hotel)" is more likely than the alternative pattern "first name (Paris), surname (Hilton) and location type (Hotel)". We argue that position measures should also used to weight the possible combinations of terms in the system to determine the more likely disambiguation. Alternatively, a rule-based approach, with hand crafted rules extracted from experimental evidences, could be combined with the current approach used in the system.

## 10.6 Conclusion

In this chapter we have presented a simplified approach to the Entity Type Disambiguation Problem based on named entity recognition in queries and we have presented an application which automatically extracts and classifies the

named entities in a query.

Named entity recognition in queries is a challenging task because many approaches proposed to address the same problem in textual documents are not effective in queries. This is because the lack of context information in short queries makes some classical named entity recognition algorithms fail. Moreover, many queries do not satisfy the natural language grammar, and orthographic and syntactic cues are often not available for disambiguation purposes.

Since local cues are scarce in queries, a reasonable approach seems to use external information to support the disambiguation process.

Traditionally, this issue has been addressed using extensive gazetteers - lists of names of people, organizations, locations, and other named entities. Indeed, the compilation of such gazetteers is sometimes mentioned as a bottleneck in the design of Named Entity recognition systems. Cucchiarelli et al. [52] report, for example, that one of the limitations in designing NE recognition systems is the limited availability of large gazetteers, particularly gazetteers for different languages. Indeed gazetteers are difficult to develop and domain sensitive. The lists need to be huge to have suitable coverage. It is estimated, for example, that there are 1.5 million unique surnames just in the USA. A gazetteer which would list all the surnames in the world should be enormous. There is a similar problem with company names. A list of all current organizations worldwide would be huge, if at all available, and would immediately be out of date since new organizations are formed all the time. In addition, organization names can occur in variations: a list of organization names might contain, for example, “Digital Enterprise Research Institute”, but that institute might also be referred to as “DERI”. The same is true for events. Consider, for example, names of conferences that usually have an extended form and an acronym. To surmount these obstacles, application of machine learning approaches (e.g. Maximum entropy, Hidden Markov Models, Memory-based Based learning) to NER became a research subject. Nevertheless all these machine learning algorithms rely on previously hand-labeled training data. Obtaining such data is labor-intensive, time consuming and usually is restricted to a specific domain.

Since our goal was to develop an application for NER in queries we needed an approach which allowed (at least potentially) to address the high heterogeneity and variability that the Web introduces. To broad coverage entity recognition we adopted a scalable approach which exploits the vast amount of RDF triples available on the Web as lists of named entities (and additional information) which have semantic annotations and extracts from these annotations attribute types and entity types. The advantages of using these annotated data are many. First of all a vast amount of RDF metadata are today available on the Web and new RDF metadata are continuously produced, facilitating the “on-line” updat-

ing and expansion of the external knowledge at the core of the disambiguation process. If a new organization is annotated, for instance, it is potentially available to be indexed in our system. Second, with the expansion of the Semantic Web, multi-language metadata will be available, allowing to address another drawback of traditional monolingual gazetteers approaches.

Another advantage of our approach compared to other approaches based on gazetteers is that while gazetteers are built as look-up lists of unique terms (useful only for perfect matches), our approach indexes terms and combinations of terms, extracted from RDF literals, which provide a sort of contextual information to be used for performing the disambiguation process. This approach can overcome another limit of the methods based on look-up lists: even if it was possible to list all possible organizations locations, people etc., there would still be the problem of overlaps between the lists. Names such as Paris, Emerson or Washington could be names of people as well as places; Philip Morris could be a person or an organization. Our approach partially resolves this problem looking at possible combinations of terms with contiguous terms which can resolve the ambiguity. For instance, when “Washington” is preceded by the term “George” our system suggests a Person instead of a Location. However, we have already noted that for other ambiguities it is needed to implement more sophisticated strategies that take into consideration the context given by other terms or attribute types within the query. In the case of “Philip Morris”, the simple co-occurrence of the two terms is not enough to eliminate the ambiguity and therefore other attribute types or entities in the query should be considered to improve the disambiguation. It is at this level that we plan to integrate experimental evidences and insights derived from our user study with the technological solution here described.

Finally, compared to the few previous studies that have tried to recognize named entities in queries [97, 59], our approach addresses scalability issues since it is based on a data structure (i.e. inverted index) which is created expressly to support the same issues in search engines. This solution has been chosen to guarantee that the system is able to handle growing amount of RDF metadata that is an essential requirement to ensure a suitable coverage of the system.

A preliminary evaluation of the system on a limited number of entity types shows that the proposed approach can accurately perform the entity disambiguation at least for two (Person and Location) out of the three types of entities which we considered. We argue that this is due to the specific composition of the data set we used to build the index, which was strongly unbalanced in favor of these two categories of entities. Nonetheless, we believe that the results show the promising potential of this approach as soon as the quality of the data set can be improved and more and more entity specific metadata will be available.

There are several issues which we plan to address in the future. Up to now we have verified the effectiveness of our method in queries in which there are only a small number of entity types. We remark that this choice was motivated by the characteristics of the data set, as well as by the availability of vocabularies from which to extract a mapping schema. We plan to extend our approach to other entity types and design a more general schema mapping to handle these types. Another topic for future work is to develop a query disambiguation algorithm which infers the type of the entity target by integrating all the information available within the query, not only the named entities but also additional information (i.e. other attribute types) that can aid the disambiguation process. The advantage of using this information has been demonstrated in our entity search experiment, as well as the impact of making the disambiguation process sensitive to the position of the information within the query. We want to implement these insights in the future evolution of our approach.



## Chapter 11

# Conclusions and Future Work

Humans construe their environment as composed of unique individuals - people, special places, pets, artworks, events - that largely represent what it is valuable and important to their own existence. We are able to identify these individuals as members of various categories (e.g.  $x$  is a politician,  $y$  is a city,  $z$  is a dog) but we are also able to uniquely identify these individuals distinguishing them from all the other members of the same category (e.g.  $x$  is Barack Obama,  $y$  is Rome and  $z$  is Fido). Every aspect of our interactions with the unique individuals relevant to our life strongly depends on our ability to correctly identify and successfully track these entities over time, change and situations. These issues seem to be a foundational component of how we perceive not only our environment but ourselves as well, by anchoring our existence to the background of our affective continuity. When these abilities go awry, the consequences can be devastating - and revealing. Consider, for example, neurological disorders such as prosopagnosia, the inability to recognize familiar faces, or the Capgras syndrome in which individuals believe that significant people in their lives have been replaced with strangers who are perceptually identical imposters.

We name singular cognition the complex of cognitive processes which allow a cognitive agent to identify a known entity, through perceptual or epistemic access to its memorial representation, and trace it as the same *unique* entity over time and change.

To perform singular cognition a cognitive agent is confronted with a uniqueness problem, i.e. the problem to identify and trace an individual as the same continuing individual, distinguishing that specific individual from all the other members of the same kind. A fundamental challenge is thus to determine how

people solve the uniqueness problem for identification and tracking of unique individuals.

Many studies have addressed the uniqueness problem in the context of visual perception (e.g. adult object-based attention) and infant cognition (e.g. object persistence and numerical identity), exploring the principles by which the visual system segments the visual input in discrete objects and bind individual views of objects into dynamic representations which persist across time, motion, featural change, and interruptions. In section 3.1 we have discussed models of vision and attention which propose direct mechanisms of individuation and reference to explain how a perceiver can perform the perceptual individuation or identification of an object as the same unique object perceived at successive moments in time (i.e. *perceptual identity*). All the models that we have reviewed share a non-conceptualist approach to the uniqueness problem in singular perception since they are based on the idea that sensory-motor capacities or perceptual contents make possible for a perceiver to latch on to, or to track a target  $x$  as being the same target without the help of complex conceptual or descriptive capacities. Such perceptual capacities must be able to perform anchoring of the perceiver onto the object  $x$  and provide perceptual reference to  $x$ , regardless of the fact that the object is fully identified as a unique individual (e.g. the object  $x$  is my dog Fido) or as a member of a category (e.g. the object  $x$  is a dog).

In this work we addressed the uniqueness problem from a different perspective, studying the conceptual system that comes into play when an object is fully identified as an instance of a conceptual representation in memory. This system comes into play when an object is identified as a known individual both in presence and in absence of perceptual information ensuring what we refer to as *conceptual identity*. Only assuming the existence of such a system we can explain how we recognize and track individual objects over long-term interruptions or changes in perceptual properties or even in absence of perceptual inputs, when the identification is based on purely descriptive information.

At the core of singular cognition we identified a system of conceptual representations, i.e. singular concepts, that are the cognitive devices specialized to uniquely identify and track individuals in different situations. Differently from singular mechanisms of reference assumed by models of visual attention (e.g. object files, FINSTs), singular concepts are conceived as long-term memory representations that allow long-term identification and entity tracking across lapses of attention, sleep, and other perceptual interruptions, as well as changes. As we have already noted in the course of this work, the representation of exemplars has to include conceptual, as well as purely perceptual, information in order to explain the way we trace identity in all the situations in which we have to track individuals on the basis of non-perceptual facts. In this sense, singu-

lar concepts represent the core of the identification process both in perceptual contexts and in contexts in which perceptual information is scarce or not available at all. Through singular concepts we are able to unite discrete glimpses or descriptions of an object into glimpses or descriptions of the same individual. We use singular concepts to keep track of properties (perceptual and conceptual) that are constant from one phase of an object to another. Identification depends, indeed, on a variety of cognitive means for information acquisition, such as perception, reasoning, communication and so on. The acknowledgment of this variety of means requires distinguishing two ways to access singular concepts: *perceptual* or *bottom-up* and *epistemic* or *top-down*. A singular concept can be accessed via a *bottom-up* way by a perceptual stimulus. In this case the individual is present in a sensory field of the agent's perceptual systems and the perceptual input activates the corresponding mental representation, through a direct match with the perceptual information stored in the concept. Alternatively, a singular concept can be accessed via a *top-down* way in cases in which the target individual cannot be perceived, but can be identified on the basis of indirect information gathered by such sources as memory, reasoning or communication. For instance, every time we talk about an absent individual, we refer linguistically to that individual by means of expressions such as singular terms or descriptions used to activate the corresponding singular concept in our interlocutor. When we think of an individual we access top-down to the mental representation (singular concept) that possesses the relevant structure of a mental device for which there is a singular content. This representation refers to, is about a single individual and it is available for purported re-identification of that individual.

In the course of the present work we have investigated different aspects of the nature and the functioning dynamics of singular concepts and we have collected the elements to sketch a model of singular cognition that has the notion of singular concept at its core. This issue has been poorly investigated in the literature on concepts in cognitive psychology that centers mainly on general concepts, such as dogs and buildings, rather than on concepts of individuals, such as a specific dog or a specific building. Moreover, studies that have addressed processes involved in singular cognition, like individual object recognition, mainly focused on a specific class of individual objects, i.e. faces - often considered a special class of individual entities - and very few studies have extended the investigation to other categories of unique individuals, such as buildings, artworks, products, organizations and events. Therefore, the first motivation of this work was to fill this gap, by proposing a model of singular cognition based on a system of mental unique representations which ensure the agent's individuation of unique objects through different contexts, time and change.

Beside the aim of better understanding some poorly investigated aspects of high-level cognition and suggesting new research directions in cognitive psychology, the present work is also motivated by a more practical need. In their interactions with information-rich spaces, such as the Web, people are more and more faced with searching and identifying individual entities in information contexts. They use search engines to find information about people, events, places, products and other entities. They access to social networks to find old friends and keep up with people with whom they contact rarely. They interact with domain information systems which store information about unique entities such as databases and digital libraries. They identify and tag people and other objects on digital pictures and so on. In all these activities people perform acts of singular cognition but they also interact with computer-based systems that face with the problem of managing entity-centric information which is in play in the course of these interactions. In this context, new approaches for addressing entity identification issues have been proposed and the idea that entities are at the core of user-information systems interactions, in particular in the case of distributed networked-based systems, has become a new frontier of investigation, being, for instance, one of the main pillar of the Semantic Web and other innovative entity-centric approaches (e.g. the Entity Named System described in Chapter 5). Therefore, the study of how people represent individual entities and how they use these representations to perform acts of identification in their interaction with information systems (e.g. to formulate keyword queries to look for specific entities) has a potential application in a technological context to improve the development of systems which are more and more designed to understand the real needs and requests of their users.

The first step of our investigation of the processes underpinning singular cognition concerned the bottom-up access to singular concepts. We explored how a perceptual (visual) stimulus makes contact with the corresponding individual-specific knowledge stored in semantic memory, that is how a perceptual stimulus makes contact with its corresponding singular concept. Traditional models of knowledge representation both in cognitive science and in computer and information sciences (see for example [3, 73]) assume a hierarchical representation of knowledge whose lower part concerns with knowledge about individuals, while the upper part concerns with knowledge about general concepts. The question we investigated was about the level where the perceptual stimulus of a unique individual first makes contact with this hierarchy (entry point). This level corresponds to the level of abstraction at which the stimulus is first identified. In other words, we explored whether the perceptual stimulus of an individual directly activates its corresponding individual representation in memory - being first recognized as a unique individual - or it first activates higher level nodes

in the hierarchy - being first recognized as a member of a category - and only after that it activates the corresponding singular concept. This issue is relevant since many studies [160, 198] have shown that although all things can be identified at different levels of abstraction, there seems to be one level, the basic level, that has a special status serving as the typical entry point in object recognition. However other studies [240] showed that individual differences in domain-specific knowledge could also influence the entry point. For example, the entry point of recognition of experts, in contrast to novices, may shift to a level that is subordinate to the basic-level. Finally, for a special kind of objects, i.e. human faces, there are evidences [241] that the entry point can be shifted to the unique level of abstraction which corresponds to the level of singular concept in our framework.

In three experiments, described in Chapter 6, we investigated the hypothesis that the entry point of unique individuals is at the level of unique identity also for other kinds of unique objects (artworks, buildings and products), indicating that this downward shift of the entry point is not a peculiarity of face stimuli but it is the common way that the cognitive system adopts to process individual entities for which the system has a singular representation in memory. Having a singular concept of an individual entity creates a preferential and direct access to that individual's specific knowledge in memory. The individual is identified as a unique individual before than being identified as a member of a basic-level category. Given that the perceptual demands increase with category specificity [119], the greatest amount of visual processing is required to identify individuals at the most specific level of unique identity. This means that in order to differentiate one familiar individual from other individuals of the same category, the recognition system must be sensitive to fine-grained differences in perceptual input. Despite these formidable constraints, our results show that the most specific category level of unique identity is the level at which individuals are first recognized when the recognition system has the cognitive devices (i.e. singular concepts) specialized to efficiently perform these kind of processing and differentiations. The entry point of individual objects which usually coincides with the basic level of classification shifts to the most subordinate level of unique identity when the perceivers become especially sensitive to subtle differences of an object compared to the other objects of the same class. We argue that this discrimination is possible thanks of the fact that a singular concept is available in memory about that specific object. Since we claim that singular concepts are initiated for those individuals we need to mark as "unique" in semantic memory, it should be more convenient for the cognitive system to identify a "unique" individual as that specific individual instead of a member of a general category. This allows, indeed, a cognitive agent to perform suitable and fast interactions

with that individual, such as performing actions or having reactions accurately directed to that specific individual.

Therefore, the first element of our model of singular cognition concerns the direct perceptual bottom-up access (i.e. non-mediated by higher level concepts) to singular concepts in semantic memory that means that the first recognition of individual entities occurs at the level of unique identity. We note that this is an assumption underlying the most important models of face recognition in literature (see for example [34, 37]) but it has remained untested for other individual objects.

As assumed by theoretical accounts of the processes underlying face recognition, we argue that the bottom-up access to the knowledge stored in the singular concept involves a match between the products of structural encoding processes of the visual stimulus and previously stored structural codes describing the appearance of familiar individuals which are part of the information stored in their corresponding singular concepts. When the match is found the corresponding singular concept is activated and the direct referential link between the mental representation and the individual in the world is established. In this way, the system has at its disposal all the information it needs to perform a suitable interaction with that specific individual.

The first phase of our investigation about the access mechanisms to singular concepts provided evidence for a direct access to unique representations in memory which is not mediated through general concepts. However, it is still open the question whether general concepts structure the semantic knowledge stored in singular concepts determining how singular concepts are interconnected each others, or whether other organization principles come into play to inter-link singular concepts. To investigate this issue we probed the semantic system using a priming experiment, described in Chapter 7.

In particular we investigated two different ways in which representations of individual entities can be related in memory. One is based on vertical relationships which connect individual instances to categories, the other is mediated by horizontal relationships between individual instances within or across categories. We name “categorical” the relationships of the first type, “associative” the relationships of the second type.

In the first case, abstracted superordinate categories are used to create a connection between individual items which belong to the same category. Two instances of the same category are connected to the representation of the category which they belong to and the category creates an indirect link between the two instances. This means that once an instance of a category is presented and recognized, activation spreads to the other instances of the same category. If semantic representations of individual entities are inter-linked by categorical

structures, we should register priming effects when prime and target entities has no other connection than the category membership. This kind of spreading activation is assumed for example by the Burton et al. [38] model of face recognition to explain how the activation of the person identity node of an individual (PIN) can activate PINs of other persons which share the same semantic categories.

The second way by which semantic representations of individual entities can be structured and connected in memory is by means of direct associative links. These links are not mediated by shared category memberships but reflect meaningful co-occurrence relationships between singular entities (which non necessarily belong to the same category). Category membership can be part of the information shared by associated singular concepts but is not the semantic connection that interlinks them. If associative links structure the representations of individual entities in memory, this means that once an entity is presented and recognized, activation spreads from the singular concept of the entity to its associated singular concepts. In terms of priming, we should obtain a priming effect when prime and target entities are associatively related even when they do not share category membership.

The results of our experiment provides support to the latter semantic organization structure of singular concepts. Once an individual which can be recognized at the level of unique identity, such as famous person, artwork, building or product, is presented and recognized, activation spreads to other individuals associatively connected to it, which produces the associative priming effects. The results of the experiment show clearly that there was no reliable categorical priming of individuals, in the sense that there was no significant benefit from primes corresponding to the proper names of members of the same category (e.g. another person from the same occupational category or another painting) but not associated. Since we found that associative links between prime and target from different categories produced facilitation effects and for the category person we found similar priming effects when associated pairs were from the same or from different categories, it appears that activation within the semantic system spreads to the representations of associates by connecting paths other than those provided by general concepts. We argue that these paths are associative in nature and may connect singular concepts of individuals from different categories, contrary to what assumed in models of face recognition.

These findings provided a second element to our model of singular cognition suggesting that semantic representations of individuals (i.e. singular concepts) are preferentially inter-connected by networks of horizontal associative links rather than by vertical categorical relations. An individual marked as “unique” in the semantic system is more strongly connected with other unique individu-

als associatively related to that individual than with other unique individuals of the same category. This suggest that the information about the category membership, such as for example the fact that Paul Newman is an actor or Mona Lisa is a portrait, may be part of the information stored in the singular concept, but it is not generally used to organize or inter-connect the network of singular representations of individuals that share the same category membership (e.g. to connect the representation of Paul Newman with that of Johnny Depp). The categorical association between unique entities may be created temporarily, as suggested by Barsalou [11] for ad hoc or goal directed categories such as “American actors”, but our results show that the representations of these entities in memory are not permanently organized by these abstracted superordinated categories.

For the specific case of person recognition, our results are in accordance with those of other studies [271, 9] which reported non-significant categorical priming in face recognition, compared with larger and statistically significant associative priming. We note that the failure to consistently observe categorical priming of person recognition challenges the Burton et al. [37] model of organization of person knowledge which assumes that the only semantic connection between person identity nodes (PINs) is through shared semantic units (SIUs) which organize biographical knowledge of people in memory creating an indirect link between PINs connected to the same SIUs. The evidence for associative priming effects suggests that associative relationship is a kind of semantic relationship that should be account by a model of person recognition. We argue that this can be done assuming some form of direct link between the identity nodes (i.e. singular concepts in our framework) of associated individuals which is reinforced when these individuals frequently co-occur. Since we found that the same associative connections can be established between persons and individual entities from different categories (e.g. between a person and an artwork), our results challenge the model proposed by Barry et al. [9] which suggest that representations of known persons are connected to each other individually by links representing specific inter-personal relatedness. We propose, instead, that significant binary properties between entities of whatever category can be transformed into an associative link if the relationship is reinforced by co-occurrence.

In conclusion, these findings show a fundamental aspect of the functioning dynamics of singular concepts and confirm what we found in the first step of our investigation: singular concepts of entities from different categories share common mechanisms of access and organization in semantic memory. Some of these mechanisms have been previously investigated in studies about person recognition. One of the main contribution of our work is to have extended these results to other kinds of individuals entities supporting the hypothesis



of the existence of a system of singular concepts in memory which is partially independent from the system of general concepts. In this sense singular concepts are not mere instances of general concepts, as assumed by many models of knowledge representations, but they are complex conceptual representations with they own functioning and organization mechanisms.

The first two phases of our work investigated how singular concepts are bottom-up accessed by perceptual (visual) stimuli and how they are interconnected each others in semantic memory. The third aspect we explored about the functioning dynamics of singular concepts deals with the organization of the information within a singular concept and the top-down access to this information in order to perform a specific identification task (e.g. search for information about specific entities by keyword queries). In our model of singular cognition, singular concepts are represented as organized structures of semantic features (or attributes) which store our knowledge about the individuals they are about. These features are of different importance in concept representations, being the most important features those that better absolve the identification function, that it the function to decide, by accessing to the system of singular concepts, if an encountered entity corresponds to an entity previously encountered and marked in memory as “unique” by means of a unique representation. In a first study we used a feature listing task paradigm to collect feature norms for individual entities from a small set of categories (person, organization, event, artifact and location) and subcategories (e.g. politician, manager, sport person, actor, professor for the category Person). The basic premise of the method was that participant’s conscious intuitions about the most important features to identify individual entities actually reflect the underlying organization of the corresponding mental representations (i.e. singular concepts) in terms of feature relevance. By collecting data from a large sample of participants, our aims was to identify patterns of attribute types that people judge more relevant to uniquely identify entities from the above mentioned categories. To this purpose we adopted a model of identification relevance, based on the model of semantic relevance proposed by Sartori and Lombardi [207] for general concepts, which computes the contribution of a feature to identify an individual of a category. The results of this analysis gave us non only an interesting overview about the distribution of attributes in terms of measures such as dominance, distinctiveness and relevance for the different categories of entities, but provided in particular an important set of data to address the second goal of the present work concerning possible contributions that a cognitive study on the identification processes in humans can provide for the development of identification algorithms in automatic systems. To this purpose, we focused on the specific case of the Entity Name System (ENS), described in 5, which provided a useful ground to explore

this issue. In particular, the results of our study have been used to address two issues concerning the functioning of the ENS: entity representation and entity matching. In the first case, we proposed to directly apply our findings to the way entities are represented in the system, suggesting a default schema for entity description to be implemented in the module of the system which manages entity entering. The second application area that we have investigated was entity matching, i.e. the process that attempts to return the single one entity that a user was (most probably) looking for when searching the ENS. We investigated two ways of how the research findings can be applied to this problem: in a straight-forward manner, to rank the results of the matching process according to the relevance of the corresponding attributes, and a “backward” manner, by making inferences about the desired type of entity from a given search term.

This latter aspect was better investigated in a second study in which we explored more specifically the search strategies used by people to formulate search queries about individual entities. On the one hand, this study gave us the opportunity to study an interesting case in which singular concepts are top-down accessed to extract a small amount of information which is used to uniquely identify an entity in the interaction with a search system. The idea of a goal-directed activation of the knowledge stored in a singular concept enriched our model of singular cognition with a new element. In particular, the results of the study integrated those obtained in the previous investigation showing which attributes people consider more relevant to identify individual entities in a specific task and the strategies they used to organize these attributes within a query.

On the other hand, the analysis of the attributes used in keyword search queries provided evidence for the beneficial impact that a cognitive study on the identification strategies used by people in entity searching may have in improving the performance of computer-based search systems. In particular, we showed how our results could contribute to address one of the most critical problems in information retrieval that is the problem to capture the meaning of a query most likely intended by the user. Our assumption was that an important first step of performing such a task is to understand what type of entity the user is looking for. We call this process Entity Type Disambiguation. To address this problem we proposed and tested a Bayesian Model based on the assumption that an entity type can be inferred from the attributes a user specifies in a search query. The beneficial impact of the entity type disambiguation approach on a search system was proved on the effect of the disambiguation on the performance of a real system. The proposed approach led finally to the development of a technological application for automatic entity type disambiguation of queries which represents a concrete example of how a cognitive investigation on the

problem of entity identification can inspire the development of technological solutions.

A further step toward the understanding of the functioning dynamics of singular concepts at the service of singular cognition concerns their role in tracking the identity of individuals in identity judgments. Because individuals can change some of their properties while persisting as the same individuals, an important aspect of singular cognition deals with the persistence of individual identity across time and change. To address this issue, singular cognition needs a function that connects different (temporal) descriptions of an individual into a unique singular description of the same individual, ensuring a unique referential link between the mental representation in the cognitive system and the corresponding referent in the world over time.

In chapter 9 we examined this fundamental aspect of singular cognition studying how people perform identity judgments among alternatives over change and we provided evidence in favor of a causal model of individual identity, i.e. Causal Continuer Model, that has the notion of causal closeness at its core. According to this model, causal continuity captures the intuition that people think of causes as central to object persistence and they use causal explanations to explain the persistence of individuals across changes. The model assumes that to decide if an individual  $x_0$  at one time  $t_0$  can be considered identical to one of a set of possible individuals  $x_1, x_2, \dots, x_n$  at another time  $t_1$ , people perform a two-step decision process. First they select a subset of candidates which are individuals that are causally close enough to be considered identical to the original and then they choose as identical the closest individual to the original among the selected candidates. In order to perform such a process we argue that the cognitive system performs a comparison between the mental representation of the original individual  $x_0$  and the representations of the possible candidates. The degree of causal closeness between the original description of  $x_0$  and the description of a candidate  $x_i$  is established evaluating how easy is to find a causal explanation that may explain the differences between the representation of  $x_0$  compared to that of  $x_i$ . In other words people decide that a singular representation about a target individual  $x_0$  at one time  $t_0$  belongs to the same object as a representation of it at another time  $t_1$ , if there is a causal link which explains the transition of the representation at  $t_0$  into the representation at  $t_1$  and this link is the strongest compared to other links which connect the original representation with representations of other possible continuers.

To test the model predictions we performed an experiment in which participants had to make identity decisions between alternative descriptions which varied in causal distance, as measured in a previous experiment, from an original description. The results of the experiment showed that the predictions of

the Causal Continuer Model fitted well the experimental results with the only exception of identity judgments about events whose ontological nature seems to determine different ways of processing.

These findings have added another “brick” to the model of singular cognition, showing that causal forces come into play in governing the mechanisms of tracking identity across time and change mediated by singular concepts. The fact that analogous results have been found for objects from different categories suggests that causal closeness is a general metric used to trace the identity of objects of different types by uniting discrete descriptions of an object into a unique singular concept of the same individual. As we have noted above, these results support the idea that system of singular concepts is subjected to common mechanisms of functioning that are not restricted to or peculiar of specific categories of objects and do not depend on the concepts or categories to which these objects belong. In this sense our work promotes a sort of revenge of singular concepts on general concepts whose predominant role in identity tracking was claimed for example by sortalist approaches to the problem of object identity. Of course, we are not deny that objects of different types can vary in their behavior in ways that are important for identity and persistence. Dropping a crystal glass from 50 cm can break it. Dropping a ball from the same height won't. But what distinguishes the causal approach with sortalist views is the explanation for such differences. While sortalist view assumes that the source of such differences is the meaning of the sortal terms that describe the objects, the causal model that we are proposing here accounts for the difference in terms of the kinds of causes responsible for maintaining the integrity of the object in question. People use their knowledge about the causal mechanisms in different domains to judge whether a certain event will cause certain consequences or not and use these knowledge to predict the compatibility of a given change in an object's description with the identity of that object.

There are a lot of other questions about singular concepts and identity that we aim to investigate in future work. We conclude this thesis by sketching possible future research directions inspired by the work reported here.

One of the most reliable findings in the literature on person identification is that semantic categorization of a face occurs more quickly than naming a face. Response latencies are slower in a name classification task (e.g. is the person's first name Paul or not?) than a semantic classification task (e.g. is the person a politician or not?). Participants are quicker to determine whether two simultaneously presented faces are those of people sharing a semantic property, for example, occupation [273] and nationality and dead/alive decisions than those of people sharing the same first name. Several explanations have been proposed for the difficulties with name retrieval relative to other biographical

information. In the Bruce and Young model [34], name retrieval takes place in a separate processing stage that follows, and is contingent upon, the retrieval of semantic information about the person. Therefore, the relative difficulty of name retrieval is the consequence of a serial architecture. In contrast, Burton and Bruce [37] proposed that names and semantic information can be accessed in parallel. The difficulty of name retrieval is a consequence of name uniqueness; whereas most semantic properties like occupation or nationality are shared by many people, names are unique to one person and therefore they take longer to be activated. The serial/parallel debate remains unresolved even though recent studies have presented evidence supporting parallel rather than serial access models [214].

To the best of our knowledge there are no mental chronometry studies that have addressed the same issue for other kinds of unique entities. Therefore, a first line of research could investigate semantic categorization compared to naming for other types of unique entity, such as buildings, artworks or products. This would be a further step to explore the parallelism in the functioning dynamics of singular concepts about people and those about unique objects from other categories.

Most published studies about person recognition and naming involve celebrities and people known through media and also the studies presented in this work used famous entities as stimuli. However, as far as recall of information (e.g. names and other semantic information) about unique individuals is concerned, the importance of frequency and recency of use may also be relevant to explain the behavioural evidence and should be considered before proposing general theoretical accounts. Frequency of exposure to names could determine the ease of name retrieval compared to recall of other information. To test this hypothesis future studies could compare semantic categorization and naming of personally known entities, such as highly familiar persons or buildings with which participants interact regularly. The results could be confronted with those obtained from famous entities of the same categories. Differences between the two conditions would support parallel access to names and semantic information in singular concepts and would show that frequency of name use could be an important determinant of data patterns in entity naming.

The experiment that we have presented in Chapter 9 investigated how people trace the identity of objects across change evaluating which changes are compatible with the identity of an individual and which are not. However, our concept of an individual object can sometimes undergo fission or fusion, even when the object itself is unchanged. As a real-life example, imagine, for example, that you never met the sister of a friend. Nonetheless, you have a singular representation about her based on what you know from him. One day you met a

girl at a party and a new singular concept is initiated about her in your memory. At a certain point during the conversation you understand that the person you have just met is in fact the sister of your friend. This means that now you have two singular representations about the same individual. In this case the two singular concepts need to be merged to create a unique representation which combine all the information stored in two original representations. The opposite process, conceptual fission, sometimes also occurs in revising our knowledge of people. For instance, going back to the previous example, if you think that your friend has only one sister, every time that he tells you about his sister you store the information into a unique mental representation. But if you discover that in fact he has two sisters, you need to revise your memorial representations and eventually distribute different pieces of information into two different concepts.

An interesting research question deals with the reorganization processes of singular concepts during these processes. In the fusion process the issue is to understand how the information initially stored in two singular representations is reorganized into a unique representation which creates a unique referential link with its referent. Anderson [3] suggests that which concept is retained and which is abandoned depends on the amount of information connected to the two. People first encode via a proposition that the two representations turned out to have the same referent. Then they choose to maintain the “stronger” representation, the one with more information. They begin a process of copying information from the abandoned representation to the other, the links to the abandoned representation are weakened through disuse and finally the access to it is lost. We argue that other mechanisms may come into play to explain which singular concept is retained and which is abandoned. In particular, the results presented in Chapter 9 suggest the hypothesis that a dimension such as the causal distance of the information contained in the two original representations may influence the revision process. In fusion cases, we might prefer to keep the concept that we can most easily “explain” becoming the merged individual who could more readily acquire the properties of the other. The idea is that the representation whose attribute changes are difficult to explain is maintained, while the other is abandoned. To test this hypothesis, we could use the Anderson’s technique<sup>1</sup> to see whether causally stable traits dominate less stable ones in conceptual fusion or fission. We could use the causal distance measures collected in the first experiment described in Chapter 9 to create the tasks and

---

<sup>1</sup>The procedure used by Anderson concerned with the speed with which subjects can retrieve facts and make inferences from them. Participants were taught with facts about different individuals and then they learned that pairs of these individuals were indeed the same individual. The latencies in making inferences from facts about these individuals were used to explain how the corresponding representations are re-organized in memory after the fusion process.

predict which of two descriptions should be retained and which should be abandoned and test the predictions on experimental data. We note that fission and fusion of singular concepts can also happen for effect of the fact that an entity, sometimes, splits into two or more entities or two entities merge in a unique entity. As a real world example, consider an organization that acquires another or two organizations that merge to form a new entity. In the context of an ENS these decisions have a big impact, since they may guide the decisions to purge entities from the repository or populate it with new entities. Therefore, understanding how these processes are managed by a cognitive system could provide interesting insights to develop algorithms to manage fusion and fission in entity-centric systems.

In summary, the contribution of this work is twofold. On one hand, we provided new evidence on the nature of high-level cognitive mechanisms involved in entity representation and identification, suggesting new research issues on a field scarcely investigated in cognitive psychology. On the other hand, we provided concrete examples of how a better understanding of these processes at a cognitive level can improve the development of entity identification approaches in information systems, suggesting a middle ground where cognitive models and technological solutions can find the opportunity for integration, in particular in information-rich spaces, like the Web, where users and machines constantly interact to satisfy entity-centric information needs.





## Appendix A

# Experimental Materials used in the Entry Point Experiments

### A.1 Entry Point Experiment 2

In Table A.1 we reported stimuli and category words used in the category-verification tasks of the Experiment 2. Each stimulus was presented six times, three times in the true condition (one for each level of abstraction of the true category word) and three times in the false condition (one for each level of abstraction of the false category word).

Stimulus	Level	Category Word	
		True Condition	False Condition
Mona Lisa	Superordinate	artwork	building
	Basic	painting	sculpture
	Subordinate	Mona Lisa	The Scream
Sunflowers	Superordinate	artwork	product
	Basic	painting	sculpture
	Subordinate	Sunflowers	The Last Supper
David	Superordinate	artwork	furnishing
	Basic	sculpture	painting
	Subordinate	David	Discobolous
Statue of Liberty	Superordinate	artwork	product
	Basic	sculpture	painting
	Subordinate	Statue of Liberty	The Pietà
Eiffel Tower	Superordinate	building	utensil
	Basic	tower	skyscraper
	Subordinate	Eiffel Tower	Leaning Tower of Pisa
Empire State Building	Superordinate	building	product
	Basic	skyscraper	church
	Subordinate	Empire State Building	Twin Towers
Golden Bridge	Superordinate	building	musical instrument

	Basic	bridge	tower
	Subordinate	Golden Bridge	Rialto Bridge
St. Peter's Basilica	Superordinate	building	product
	Basic	church	bridge
	Subordinate	St. Peter's Basilica	Milan Cathedral
Fiat 500	Superordinate	product	building
	Basic	car	audio player
	Subordinate	Fiat 500	Mini Cooper
Fiat Panda	Superordinate	product	artwork
	Basic	car	phone
	Subordinate	Fiat Panda	Beetle
Iphone	Superordinate	product	artwork
	Basic	phone	car
	Subordinate	Iphone	Black Barry
Ipod nano	Superordinate	product	furnishing
	Basic	audio player	car
	Subordinate	Ipod nano	Walkman
rocking chair	Superordinate	furnishing	artwork
	Basic	chair	table
	Subordinate	rocking chair	folding chair
desk lamp	Superordinate	furnishing	musical instrument
	Basic	lamp	table
	Subordinate	desk lamp	floor lamp
tea table	Superordinate	furnishing	utensil
	Basic	table	lamp
	Subordinate	tea table	dinning table
four poster bed	Superordinate	furnishing	musical instrument
	Basic	bed	chair
	Subordinate	four poster bed	cot
wooden spoon	Superordinate	utensil	musical instrument
	Basic	spoon	pan
	Subordinate	wooden spoon	teaspoon
bread knife	Superordinate	utensil	artwork
	Basic	knife	spoon
	Subordinate	bread knife	flick knife
fry pan	Superordinate	utensil	building
	Basic	pan	knife
	Subordinate	fry pan	saucepan
nail scissors	Superordinate	utensil	furnishing
	Basic	scissors	pan
	Subordinate	nail scissors	garden scissors
grand piano	Superordinate	musical instrument	utensil
	Basic	piano	trumpet
	Subordinate	grand piano	upright piano
bongo drum	Superordinate	musical instrument	furnishing
	Basic	drum	guitar
	Subordinate	bongo drum	bass drum
trombone	Superordinate	musical instrument	utensil
	Basic	trumpet	piano
	Subordinate	trombone	clarinet
electric guitar	Superordinate	musical instrument	building
	Basic	guitar	drum
	Subordinate	electric guitar	acoustic guitar

Table A.1: Stimuli and categories words used in experiment 2

## A.2 Entry Point Experiment 3

In Table A.2 we reported stimuli and word primes used in the identity-matching tasks of the Experiment 3. In the experiment, a word prime was followed by two pictures. In the same condition two identical pictures were presented. In the different condition the stimulus picture was paired with a picture of a different object. In the Table A.2 we reported the list of the stimuli pictures for the same and different conditions and the prime words used for the three levels of abstractions (i.e. basic, subordinate and neutral).

Stimulus	Level	Category Word (prime)	Paired with (different condition)
The Scream	Basic Level	painting	Sunflowers
	Subordinate Level	The Scream	Mona Lisa
	Neutral	blank	The Kiss
The Last Supper	Basic Level	painting	The luncheon of the boating party
	Subordinate Level	The Last Supper	Starry Night
	Neutral	blank	The Birth of Venus
Discobolous	Basic Level	sculpture	Statue of Liberty
	Subordinate Level	Discobolous	David
	Neutral	blank	Venus de Milo
The Pietà	Basic Level	sculpture	Riace Bronzes
	Subordinate Level	The Pietà	The Thinker
	Neutral	blank	The Kiss
The Leaning Tower of Pisa	Basic Level	tower	Eiffel Tower
	Subordinate Level	The Leaning Tower of Pisa	Big Ban
	Neutral	blank	Asinelli Tower
Twin Towers	Basic Level	skyscraper	Empire State Building
	Subordinate Level	Twin Towers	Taipei 101
	Neutral	blank	Sears Tower
Rialto Bridge	Basic Level	bridge	Golden Gate Bridge
	Subordinate Level	Rialto Bridge	Old Bridge
	Neutral	blank	Tower Bridge
Milan Cathedral	Basic Level	church	Cathedral of Notre-Dame
	Subordinate Level	Milan Cathedral	St. Peter's Basilica
	Neutral	blank	Canterbury Cathedral
Mini	Basic Level	car	Lancia Y
	Subordinate Level	Mini Cooper	Fiat 500
	Neutral	blank	Smart
Beetle	Basic Level	car	Golf
	Subordinate Level	Beetle	Fiat Panda
	Neutral	blank	Peugeot 206
Black Barry	Basic Level	phone	Iphone
	Subordinate Level	Black Barry	Nokia E71
	Neutral	blank	Samsung I8000
Walkman	Basic Level	audio player	Ipod Nano
	Subordinate Level	Walkman	Ipod shuffle
	Neutral	blank	Ipod classic
folding chair	Basic Level	chair	rocking chair
	Subordinate Level	folding chair	office chair
	Neutral	blank	armchair
floor lamp	Basic Level	lamp	desk lamp
	Subordinate Level	floor lamp	night table lamp

	Neutral	blank	table lamp
dinning table	Basic Level	table	tea table
	Subordinate Level	dinning table	picnic table
	Neutral	blank	billiards table
cot	Basic Level	bed	four poster bed
	Subordinate Level	cot	cradle
	Neutral	blank	iron bed
tea spoon	Basic Level	spoon	wooden spoon
	Subordinate Level	tea spoon	Chinese spoon
	Neutral	blank	honey dipper
flick knife	Basic Level	knife	cheese knife
	Subordinate Level	flick knife	bread knife
	Neutral	blank	kitchen knife
saucepan	Basic Level	pan	fry pan
	Subordinate Level	saucepan	pasta pan
	Neutral	blank	pressure cooker
garden scissors	Basic Level	scissors	nail scissors
	Subordinate Level	garden scissors	chicken scissors
	Neutral	blank	barber scissors
upright piano	Basic Level	piano	grand piano
	Subordinate Level	upright piano	spinet
	Neutral	blank	keyboard
bass drum	Basic Level	drum	bongo drum
	Subordinate Level	bass drum	tambourine
	Neutral	blank	congas
transverse flute	Basic Level	flute	recorder
	Subordinate Level	transverse flute	piccolo
	Neutral	blank	penny whistle
acoustic guitar	Basic Level	guitar	electric guitar
	Subordinate Level	acoustic guitar	bass guitar
	Neutral	blank	resonator guitar

Table A.2: Stimuli and categories words used in experiment 3

## Appendix B

# Experimental Materials used in the Entity Recognition Experiment

### B.1 Entity Recognition Experiment

The table B.1 shows target stimuli and word primes used in the entity recognition experiment. Numbers in the first column show the degree of association, measured as the proportions of participants ( $n=15$ ), who gave the name as the “first that springs to mind” when presented with the target name in the pilot study.

Person: across			
Associate prime	Categorical prime	Unrelated prime	Target
<b>Set A</b>			
USA (0.80)	Nicolas Sarkozy	Florence	Barack Obama
Pretty Woman (0.60)	Monica Bellucci	Nokia	Julia Roberts
Beatles (0.80)	Freddy Mercury	Pantheon	John Lennon
Ferrari (0.93)	Ayrton Senna	Poland	Michael Schumacher
<b>Set B</b>			
Mediaset (0.60)	Angela Merkel	Taiwan	Silvio Berlusconi
Mission	Robert De Niro	Chinese Wall	Tom Cruise
Impossible (0.60)			
Argentina (1)	David Beckham	Times Square	Diego A. Maradona
Albachiara (0.46)	Tiziano Ferro	Chicago	Vasco Rossi
<b>Set C</b>			
Saturday Night	Leonardo Di Caprio	Egypt	John Travolta
Fever (0.47)			
England (0.66)	Princess Grace	Panasonic	Lady Diana
Thriller (0.53)	Madonna	Trevi Fountain	Micheal Jackson
Yamaha (0.60)	Marco Melandri	Broadway	Valentino Rossi
Person: within			
<b>Set A</b>			
Angelina Jolie (0.86)	Johnny Depp	Luciana Litizzetto	Brad Pitt
Romina Power (0.93)	Andrea Bocelli	Hugh Grant	Albano Carrisi
Sandra Mondaini (0.93)	Pippo Baudo	Bob Marley	Raimondo Vianello
Ilary Blasi (0.60)	Rino Gattuso	Kelly Minogue	Francesco Totti
<b>Set B</b>			
Katia Ricciarelli (0.53)	Paolo Bonolis	Penelope Cruz	Pippo Baudo
Enzo Iacchetti (0.60)	Claudio Bisio	Joaquin Cortez	Ezio Greggio

John F. Kennedy (0.60)	Audrey Hepburn	Winston Churchill	Marilyn Monroe
Fidel Castro (0.33)	Nelson Mandela	Kevin Costner	Ernesto Che Guevara
<b>Set C</b>			
Gino Paoli (0.66)	Gianna Nannini	Cary Grant	Ornella Vanoni
Monica Lewinsky (0.60)	Tony Blair	Sean Connery	Bill Clinton
Claudio Bisio (0.56)	Daria Bignardi	Lucio Dalla	Vanessa Encontrada
Carlo Conti (0.46)	Elizabeth Taylor	Dino Zof	Sophia Loren
<b>Artwork</b>			
<b>Set A</b>			
Louvre (1)	Guernica	Vladimir Putin	Mona Lisa
Van Gogh (0.86)	The School of Athens	Cuba	The Sunflowers
Munch (0.80)	The Three Graces	Sanghai	The Scream
Greece (0.33)	The Venus de Milo	Cindy Lauper	Riace Bronzes
<b>Set B</b>			
Michelangelo (0.86)	Discobolus	Elvis Presley	La Pietà
Leonardo Da Vinci (0.86)	Water-Lilies	Barilla	Last Supper
Sistine Chapel (0.40)	Dead Christ	Bombay	The Creation of Adam
Gustav Klimt (0.33)	Luncheon Of The Boating Party	Australia	The Kiss
<b>Set C</b>			
USA (0.80)	Christ Redeemer	Switzerland	Statue of Liberty
Botticelli (0.66)	The Tree of Life	Los Angeles	The Birth of Venus
Van Gogh (0.46)	Girl with a Pearl Earring	Nikon	Starry Night
Florence (0.73)	The Thinker	Red Square	David
<b>Building</b>			
<b>Set A</b>			
Paris (0.93)	Leaning Tower of Pisa	Moscow	Eiffel Tower
Rome (0.93)	Triumphal Arc	Ibiza	Colosseum
Barcelona (0.73)	Santa Maria Novella	Ariston	Sagrada Familia
Washington (0.80)	Villa of Arcore	Luigi Pirandello	White House
<b>Set B</b>			
Bin Laden (0.86)	Tower of London	Napoleone Bonaparte	Twin Towers
Queen Elisabeth (0.86)	Palace of Versailles	Turin	Buckingham Palace
Rome (0.93)	Notre Dame	Japan	Basilica of Saint Peter
New York (0.86)	Golden Gate Bridge	The Great Pyramid of Giza	Brooklyn Bridge
<b>Set C</b>			
London (1)	Asinelli Tower	Spain	Big Ben
Berlin (0.66)	The Arch of Constantine	Toronto	Brandenburg Door
New York (0.73)	Taipei Financial Center	Michelle Hunziker	Empire State Building
Paris (0.93)	The Uffizi Gallery	Microsoft	Louvre
<b>Product</b>			
<b>Set A</b>			
Apple (0.86)	Black Barry	Martin Scorsese	Iphone
Fiat (1)	Micra	Dublin	Panda
Ferrero (0.60)	Kit Kat	Prague	Nutella
Volkswagen (0.73)	Peugeot 205	Nivea	Golf
<b>Set B</b>			
Apple (0.73)	Walkman	Portugal	Ipod
Fiat (0.86)	Ypsilon	Vienna	500
Algida (0.80)	Maxi Bon	Acer	Cornetto
Piaggio (0.60)	Monster	Trafalgar Square	Vespa
<b>Set C</b>			
Sony (0.60)	Xbox	Vatican Museums	Play Station
Volkswagen (0.73)	Megane	Kensington Gardens	Beetle
Ford (0.80)	Punto	Munich	Fiesta
Piaggio (0.53)	Scarabeo	Albert Einstein	Ciao

Table B.1: Prime words and Stimuli used in the entity recognition experiment

## Appendix C

# Relevance Measures

### C.1 Feature Norms for Individual Entities

Category	Attributes (it)	F	f	Attributes (eng)	F	f
<i>Politician</i>	age	17	0.56	party	24	0.76
	name	14	0.46	name	19	0.63
	political view	14	0.46	age	13	0.43
	party	13	0.43	country	10	0.33
	surname	11	0.43	gender	10	0.26
	type	11	0.36	role	8	0.26
	role	10	0.36	nationality	5	0.13
	education	9	0.30	surname	5	0.13
	experiences	7	0.30			
	curriculum	5	0.23			
		N=30		N=30		
<i>Manager</i>	name	13	0.46	name	0.71	16
	surname	11	0.39	age	0.28	7
	company	8	0.28	department	0.23	5
	age	7	0.25	experience	0.20	5
	role	7	0.21			
	type	6	0.21			
	education	6	0.21			
	N=28		N=21			
<i>Professor</i>	name	13	0.52	name	0.87	21
	specialization	16	0.64	university	0.41	10
	age	9	0.36	department	0.33	8
	surname	8	0.32	education	0.29	7
	educational institution	6	0.24	publication	0.29	7
	publications	5	0.20	age	0.20	5
	type	5	0.20	email	0.20	5
				research area	0.20	5
				surname	0.20	5
		N=25		N=24		
<i>Sportsperson</i>	type of sport	20	0.66	name	0.63	19
	age	14	0.46	type of sport	0.5	18
	name	14	0.46	age	0.33	10
	surname	9	0.23	gender	0.26	9
	type	7	0.23	birth-date	0.23	7
	birth date	6	0.20	nationality	0.16	5
	level	6	0.20	team	0.16	5
	N=30		N=26			
<i>Actor/actress</i>	age	16	0.51	name	0.88	16
	type	16	0.51	birth date	0.38	7
	name	15	0.48	movies	0.38	7
	experiences	14	0.45	gender	0.33	6
	nationality	11	0.35	country	0.27	5
	surname	10	0.32	age	0.22	4
	movies	10	0.32			
	birth date	7	0.22			
	N=31		N=18			
<i>Person</i> neutral category	name	20	0.74	name	0.73	19
	surname	17	0.62	gender	0.46	14
	birth-date	10	0.37	birth-date	0.42	11
	age	10	0.37	age	0.38	10
	birth-place	8	0.37	education	0.23	6
	tax code	8	0.29	height	0.23	6
	occupation	7	0.29	nationality	0.23	6
	height	7	0.25	occupation	0.23	6
	place of residence	7	0.25	surname	0.23	6
	type	7	0.25	birth-place	0.19	5
	character	6	0.22	email	0.19	5
	weight	6	0.22	marital status	0.15	4
		N=27		N=26		

Table C.1: Features and production frequencies for PERSON



Category	Attributes (all)	F	f	
<i>Politician</i>	party	37	0.61	
	name	33	0.55	
	age	30	0.50	
	role	23	0.38	
	experiences - career	19	0.31	
	political view	17	0.28	
	surname	17	0.28	
	education	13	0.21	
	country	11	0.18	
	type	11	0.18	
	gender	10	0.16	
		N=60		
	<i>Manager</i>	name	29	0.59
age		14	0.28	
role		12	0.24	
company		11	0.22	
experiences		10	0.20	
education		9	0.18	
competence		9	0.18	
		N=49		
<i>Professor</i>	name	34	0.69	
	specialization	20	0.40	
	age	14	0.28	
	surname	13	0.26	
	publications	12	0.24	
	university/ies	11	0.22	
	department	10	0.20	
		N=49		
<i>Sportsperson</i>	type of sport - specialty	38	0.63	
	name	33	0.55	
	age	24	0.40	
	birth date	13	0.21	
	gender	9	0.15	
	surname	9	0.15	
		N=60		
<i>Actor/actress</i>	name	31	0.63	
	age	20	0.40	
	type	18	0.36	
	movies	17	0.34	
	birth date	14	0.28	
	experiences	14	0.28	
	nationality	13	0.26	
	education	8	0.16	
		N=49		
<i>Person</i> neutral category	name	39	0.73	
	surname	23	0.43	
	birth date	21	0.39	
	age	20	0.37	
	birth place	15	0.28	
	gender	14	0.26	
	occupation	14	0.26	
	height	13	0.24	
	nationality	11	0.20	
	eyes color	8	0.15	
		N=53		

Table C.2: Features and production frequencies for PERSON: aggregated data (i.e. English and Italian).

Subcategory	Attributes (it)	F	f	Attributes (eng)	F	f
<i>Company</i>	name	22	0.68	name	15	0.83
	location	11	0.31	address	7	0.38
	type	10	0.31	location	7	0.38
	num. of employees	9	0.28	country	6	0.33
	turnover	6	0.18	business type	4	0.22
	sector	5	0.15	num. of employees	4	0.22
				web site url	4	0.22
		N=32		N=18		
<i>Association</i>	name	17	0.60	name	13	0.52
	objective/s	16	0.57	objective/s	10	0.40
	type	11	0.39	location	8	0.32
	members	8	0.28	type	6	0.24
	sector	6	0.21	website url	6	0.24
	location	5	0.17	activity	5	0.20
	headquarters	5	0.17	address	5	0.20
				date of foundation	5	0.20
		N=28	members	5	0.20	
				N=25		
<i>University</i>	location	15	0.48	name	16	0.61
	name	14	0.45	location	12	0.46
	faculties	9	0.25	address	8	0.30
	courses	7	0.22	city	7	0.26
	city	6	0.19	number of students	7	0.26
	num. of students	5	0.16	country	6	0.23
				courses	5	0.19
			faculties	5	0.19	
		N=31	state	5	0.19	
				N=26		
<i>Government</i>	political orientation	8	0.34	country	15	0.55
	nation	7	0.30	name	7	0.25
	type	6	0.26			
	country	4	0.17			
		N=23		N=27		
<i>Agency</i>	name	16	0.65	name	13	0.65
	type	13	0.54	address	7	0.35
	location	7	0.29	num. of employees	5	0.25
	address	6	0.25	type	5	0.25
	objective/s	6	0.25			
	num. of employees	5	0.20			
	sector	4	0.16			
		N=24		N=20		
<i>Organization</i>	name	17	0.56	name	13	0.61
	neutral category	17	0.56	location	6	0.28
	objective/s	10	0.33	type	6	0.28
	type	8	0.26			
	sector	8	0.26			
	location	7	0.20			
	head office	6	0.20			
members	5	0.16				
		N=30		N=21		

Table C.3: Features and production frequencies for ORGANIZATION

<b>Subcategory</b>	<b>Attributes (all)</b>	<b>F</b>	<b>f</b>
<i>Company</i>	name	37	0.74
	location	18	0.36
	number of employees	13	0.26
	business type	12	0.24
	address	10	0.20
	turnover	9	0.18
	country	8	0.16
	N=50		
<i>Association</i>	name	30	0.56
	objective/s	26	0.49
	type	17	0.32
	location	13	0.24
	activity	9	0.17
	N=53		
<i>University</i>	name	30	0.52
	location	27	0.47
	faculties	14	0.24
	city	13	0.23
	number of students	12	0.21
	courses	12	0.21
	address	11	0.19
	N=57		
<i>Government</i>	country	26	0.52
	type	9	0.18
	political view	8	0.16
	N=50		
<i>Agency</i>	name	29	0.65
	type	18	0.40
	address	13	0.29
	number of employee	10	0.23
	location	10	0.23
	N=44		
<i>Organization</i> neutral category	name	30	0.58
	type	16	0.31
	location	13	0.25
	sector	8	0.15
	N=51		

Table C.4: Features and production frequencies for ORGANIZATION: aggregated data (i.e. English and Italian)

Subcategory	Attributes (it)	F	f	Attributes (eng)	F	f	
<i>Conference</i>	location	18	0.9	location	22	0.84	
	topic/s	14	0.7	name	15	0.57	
	participants	10	0.5	date	10	0.34	
	date/s	9	0.45	organizer/s	7	0.26	
	duration	6	0.30	participants	5	0.19	
	speaker/s	6	0.30	topic/s	5	0.19	
	title	5	0.25	year	5	0.19	
	organizers	4	0.20				
	objective/s	4	0.20				
	time	4	0.20				
		N=20			N=26		
	<i>Meeting</i>	location	16	0.84	location	22	0.88
		time	12	0.63	time	20	0.80
topic/s		11	0.57	date	16	0.64	
participants		10	0.52	participants	13	0.52	
date		9	0.21	topic/s	9	0.35	
type		4	0.21	type	6	0.24	
				name	5	0.20	
	N=19			N=25			
<i>Exhibition</i>	location	16	0.72	location	12	0.75	
	topic/s	12	0.54	name	8	0.5	
	title	10	0.45	time	6	0.37	
	date	7	0.31	date	6	0.37	
	duration	7	0.31	end date	3	0.18	
	type	6	0.27	start date	3	0.18	
	artists	4	0.18				
		N=22			N=16		
<i>Show</i>	type	28	0.75	name/title	11	0.64	
	location	27	0.73	location	9	0.52	
	date	16	0.43	actors	5	0.30	
	duration	11	0.29	time	5	0.30	
	time	11	0.29	type	4	0.23	
	price/s	10	0.27	date	4	0.23	
	title	9	0.24				
	actors	8	0.21				
	participants	6	0.16				
		N=37			N=17		
	<i>Sport event</i> neutral category	location	16	0.76	location	17	1
date		11	0.52	type of sport	11	0.64	
sport specialty		10	0.47	date	10	0.58	
name		8	0.38	time	5	0.29	
type		6	0.28	duration	4	0.23	
time		4	0.19	name	4	0.23	
participants		4	0.19	participants	4	0.23	
	N=21			N=17			
<i>Event</i> neutral category	date	15	0.71	location	21	0.91	
	location	15	0.71	date	13	0.56	
	type	14	0.66	time	12	0.52	
	participants	8	0.38	participants	8	0.34	
	name	7	0.33	name	7	0.30	
	duration	5	0.23	type	6	0.26	
	organizer/s	5	0.23	purpose/s	5	0.21	
	time	5	0.23				
	N=21			N=23			

Table C.5: Features and production frequencies for EVENT

<b>Subcategory</b>	<b>Attributes (all)</b>	<b>F</b>	
<i>Conference</i>	location	40	0.86
	title/name	20	0.43
	topic/s	19	0.41
	date/s	19	0.41
	participants	15	0.33
	organizers	11	0.24
	objective/s	8	0.17
	duration	8	0.17
	N=46		
<i>Meeting</i>	location	38	0.86
	time	32	0.72
	date	25	0.56
	participants	23	0.52
	topic/s	20	0.45
	type	10	0.22
	N=19		
<i>Exhibition</i>	location	28	0.73
	name	18	0.47
	date	13	0.34
	time	9	0.23
	duration	9	0.23
	type	7	0.18
	N=38		
<i>Show</i>	location	36	0.66
	type	32	0.59
	title/name	20	
	date	20	0.37
	time	16	0.29
	actor/s	13	
	duration	12	0.22
	price	11	0.20
	participants	9	0.16
	N=54		
<i>Sport event</i>	location	33	0.86
	date	21	0.55
	sport specialty	21	0.55
	name	12	0.31
	time	9	0.23
	type	9	0.23
	participants	8	0.21
	duration	6	0.16
	N=38		
<i>Event</i> neutral category	location	36	0.81
	date	28	0.63
	type	20	0.45
	time	17	0.38
	participants	16	
	name	14	0.31
	duration	8	0.18
	organizers	8	0.18
	purpose/s	7	0.16
	N=44		

Table C.6: Features and production frequencies for EVENT: aggregated data (i.e. English and Italian)

Subcategory	Attributes (it)	F	f	Attributes (eng)	F	f
<i>Product</i>	price/s	15	0.6	price/s	23	0.52
	name	11	0.44	name	9	0.39
	use	11	0.44	color	5	0.21
	type	10	0.40	description	5	0.21
	color/s	9	0.32	size	4	0.17
	dimension/s	7	0.28	use	4	0.17
	features	5	0.20			
	weight	5	0.20			
		N=25			N=23	
<i>Artwork</i>	artist/s	27	0.9	creation date	14	0.73
	title/name	11	0.36	artist/s	13	0.68
	date	13	0.43	title/s	13	0.68
	location	12	0.40	material	6	0.31
	type	10	0.33	style	6	0.31
	material	9	0.30	type	5	0.26
	style	8	0.26	date	4	0.21
	color/s	6	0.20			
	subject	6	0.20			
	creation date	5	0.16			
	size	5	0.16			
		N=30			N=19	
<i>Building</i>	location	16	0.53	address	15	0.65
	height	11	0.36	location	13	0.56
	number of floors	11	0.36	height	11	0.36
	color/s	10	0.33	name	8	0.47
	dimension/s	10	0.33	architect	6	0.34
	type	10	0.33	color	5	0.26
	address	7	0.30	number of floors	5	0.21
	recipients	6	0.20	owner	5	0.21
	area mq	6	0.20	type	0.21	
	use	6	0.20	country	0.17	
	N=30			N=23		
<i>Book</i>	author/s	27	0.90	author/s	20	0.8
	title	22	0.73	title	19	0.76
	publisher	18	0.6	ISBN	13	0.52
	number of pages	16	0.53	publisher	13	0.52
	year of publication	12	0.40	year of publication	13	0.52
	type	9	0.30	number of pages	7	0.28
	ISBN	7	0.23	year	5	0.20
	topic	6	0.20	language/s	4	0.16
	edition	6	0.20	topic	4	0.16
	N=30			N=25		
<i>Article of clothing</i>	color/s	27	0.83	color	18	0.72
	size/s	21	0.68	size	13	0.52
	type	19	0.61	type	11	0.44
	brand name	15	0.48	material	9	0.36
	price/s	12	0.38	gender	7	0.28
	fabric	8	0.25	price/s	7	0.28
	material	5	0.16	style	7	0.28
	model	5	0.16	brand name	5	0.20
	N=31			N=25		
<i>Object</i>	color/s	20	0.91	size	16	0.64
	function/use	20	0.91	color	14	0.56
	material	14	0.63	shape	10	0.40
	shape	14	0.63	function	10	0.40
	size	14	0.63	name	6	0.24
	weight	10	0.45	dimensions	4	0.16
	name	9	0.40	material	4	0.16
				weight	4	0.16
	N=22			N=25		

Table C.7: Features and production frequencies for ARTIFACT

<b>Subcategory</b>	<b>Attributes (all)</b>	<b>F</b>	<b>f</b>
<i>Product</i>	price/s	27	0.56
	name	20	0.41
	color	14	0.29
	type	13	0.27
	manufacturer	11	0.23
	size	11	0.23
	description	8	0.16
	use/function	7	0.15
	features	7	0.15
	brand	7	0.15
		N=48	
<i>Artwork</i>	artist/s	40	0.82
	title/name	27	0.49
	creation date	23	0.47
	material	15	0.30
	type	15	0.30
	style	14	0.29
	location	0.29	
	size	9	0.18
	color	8	0.16
		N=49	
<i>Building</i>	location	29	0.54
	address	22	0.41
	height	22	0.41
	number of floors	16	0.30
	color	15	0.28
	type	15	0.28
	size	13	0.24
	architect	9	0.17
	name	8	0.15
	use	6	0.20
		N=53	
<i>Book</i>	author/s	47	0.85
	title	41	0.74
	publisher	31	0.56
	year of publication	25	0.45
	number of pages	23	0.42
	ISBN	20	0.36
	type	11	0.20
	topic	10	0.18
	edition	9	0.18
		N=55	
<i>Article of clothing</i>	color	45	0.80
	size	34	0.61
	type	30	0.53
	brand name	20	0.36
	price	19	0.34
	material	14	0.25
	fabric	12	0.21
	style	11	0.19
	gender intended for	9	0.16
	N=56		
<i>Object</i> neutral category	color/s	34	0.72
	size	30	0.63
	function	30	0.63
	shape	24	0.51
	material	18	0.38
	name	15	0.31
	weight	14	0.29
	use	6	0.27
	N=47		

Table C.8: Features and production frequencies for ARTIFACT: aggregated data (i.e. English and Italian)

Subcategory	Attributes (it)	F	f	Attributes (eng)	F	f
<i>Tourist Location</i>	location	14	0.58	name	11	0.57
	attractions	8	0.33	country	8	0.42
	name	7	0.29	geo. position	7	0.37
	type	7	0.29	city	5	0.26
	population	4	0.16	attractions	5	0.21
	geo. position	4	0.16	location	4	0.21
	services	4	0.16	price/s	4	0.21
		N=24			N= 19	
<i>City</i>	number of citizens	24	0.8	country	13	0.52
	country	18	0.60	name	13	0.52
	geo. position	18	0.60	population	11	0.44
	name	12	0.40	location	10	0.40
	region	10	0.33	geo. position	7	0.28
	climate	6	0.20	language/s	5	0.20
		N=30		num. of citizens	4	0.16
				N= 25		
<i>Shop</i>	location	20	0.55	location	0.38	
	name	19	0.52	name	0.38	
	type	14	0.33	height	0.47	
	address	12	0.33	address	0.33	
	timetable	11	0.30			
	number of employee	7	0.19			
	dimensions	6	0.16			
	N=36			N=21		
<i>Hotel</i>	location	18	0.66	name	20	0.83
	name	15	0.55	address	12	0.5
	services	11	0.40	location	9	0.37
	number of rooms	10	0.37	country	7	0.29
	number of stars	10	0.37	city	7	0.29
	category	8	0.29	number of rooms	4	0.16
	address	6	0.22	rating	4	0.16
	price/s	5	0.18	state	4	0.16
		N=27			N=24	
<i>Restaurant</i>	type	21	0.7	name	21	0.75
	location	17	0.56	address	20	0.71
	name	13	0.43	type of cuisine	19	0.67
	price/s	10	0.33	location	13	0.46
	address	9	0.30	price/s	11	0.39
	timetable	7	0.23	city	8	0.28
	category	5	0.16	country	6	0.21
		N=30		chef	5	0.17
			type	5	0.17	
				N=28		
<i>Place</i>	geo. position	16	0.72	geo. position	26	0.92
	name	7	0.32	country	14	0.5
	location	6	0.27	name	13	0.46
	altitude	4	0.18	city	10	0.36
	region	4	0.18	address	5	0.17
	type	4	0.18	state	5	0.17
	N=22			N=28		

Table C.9: Features and production frequencies for LOCATION



<b>Subcategory</b>	<b>Attributes (all)</b>	<b>F</b>	<b>f</b>
<i>Tourist Location</i>	name	18	0.42
	location	18	0.42
	geographical position	14	0.33
	attractions	12	0.27
	country	4	0.25
	type	9	0.20
	city	7	0.16
	price/s	7	0.16
	N=43		
<i>City</i>	country	31	0.56
	name	25	0.45
	geographical position	25	0.45
	population	15	0.27
	location	13	0.24
	N=55		
<i>Shop</i>	location	29	0.51
	name	27	0.47
	type	24	0.33
	address	19	0.33
	type of products	18	
	timetable	11	0.19
	N=57		
<i>Hotel</i>	name	35	0.68
	location	27	0.53
	address	18	0.35
	number of rooms	14	0.27
	services	13	0.25
	number of stars	13	0.26
	city	11	0.22
	category	9	0.17
	country	8	0.15
	price/s	8	0.15
	N=51		
<i>Restaurant</i>	name	34	0.58
	location	30	0.52
	address	29	0.5
	type	26	0.44
	price/s	21	0.36
	type of cuisine	19	0.32
	timetable	9	0.16
	N=58		
<i>Place</i> neutral category	geographical position	42	0.84
	country	20	0.40
	name	20	0.40
	city	11	0.22
	location	9	0.18
	N=50		

Table C.10: Features and production frequencies for LOCATION: aggregated data (i.e. English and Italian)

PERSON				
Category	Attributes (eng)	<i>k</i>	Attributes (it)	<i>k</i>
<i>Politician</i>	party	65.78	party	35.63
	name	31.20	political view	28.67
	gender	16.42	name	22.99
	position/s	14.60	age	21.02
	age	11.30	surname	18.06
<i>Manager</i>	name	26.27	company	21.93
	experiences	11.68	name	21.34
	role/s	10.30	surname	18.06
	department/s	9.12	experiences	9.12
	occupation	7.01	education	9.85
<i>Professor</i>	university	37.77	specialization	54.94
	name	34.48	name	21.34
	publications	19.19	institution	20.60
	research area	17.17	publications	17.17
	department	14.60	surname	13.14
<i>Sportsperson</i>	type of sport	49.34	type of sport	54.82
	name	31.20	name	22.99
	team	17.17	age	17.31
	birth-date	11.50	surname	11.50
	gender	14.78	birth-date	9.85
<i>Actor</i>	name	26.27	movies	34.34
	movies	19.19	name	24.63
	birth-date	11.50	experiences	25.54
	gender	9.85	age	19.79
	awards	6.14	surname	16.42
<i>Person</i>	name	31.20	name	32.84
	gender	22.99	surname	27.92
	birth-date	18.06	birth-place	18.25
	occupation	14.01	occupation	18.68
	religion	13.74	birth-date	16.42

Table C.11: Relevance Measure for PERSON

ORGANIZATION				
Category	Attributes (eng)	<i>k</i>	Attributes (it)	<i>k</i>
<i>Company</i>	name	24.63	name	36.12
	ceo name	8.22	number of employees	14.78
	business type	8.19	turnover	12.29
	profits	7.01	share capital	10.96
	revenue	6.87	activity	10.30
<i>Association</i>	name	21.34	name	27.91
	objective/s	16.42	members	27.47
	members	11.68	objective/s	11.61
	activity	11.68	number of members	7.01
	date of foundation	9.12	functions	6.87
<i>University</i>	name	26.27	faculties	30.91
	number of students	19.19	name	22.99
	faculty/ies	16.35	number of students	17.17
	courses	13.70	courses	16.35
	department/s	9.12	professors	13.74
<i>Government</i>	name	11.49	political view	16.38
	head	9.34	duration	13.74
	members	9.34	party/s	10.96
	party	8.22	ministries	10.30
	leaders	8.22	ministers	10.30
<i>Agency</i>	name	21.34	name	26.27
	number of employees	7.44	number of employees	8.21
	president	6.87	clients	6.87
	specialization	4.67	sector	4.53
	profit/s	4.67	objective/s	4.36
<i>Organization</i>	name	21.34	name	27.91
	business type	6.14	objective/s	12.34
	objective/s	4.93	members	11.68
	character/s	4.67	sector	9.05
	head	4.67	date of foundation	8.21

Table C.12: Relevance for ORGANIZATION

EVENT				
Category	Attributes (eng)	<i>k</i>	Attributes (it)	<i>k</i>
<i>Conference</i>	name	22.32	topic	25.54
	organizers	12.77	speakers	16.45
	date	12.37	participants	16.42
	chair/s	10.30	date	11.13
	sessions	10.30	needs	10.30
<i>Meeting</i>	time	29.76	topic	20.07
	date	19.79	time	17.86
	topic/s	18.43	participants	16.42
	participants	16.08	date	11.13
	agenda	13.74	location	7.01
<i>Exhibition</i>	name	11.90	topic	21.89
	time	8.93	duration	11.50
	date	7.42	artists	10.96
	start date	7.01	exhibitors	10.30
	end date	4.93	title	9.85
<i>Show</i>	actors	17.17	date	19.79
	name	13.39	actors	18.68
	producer/s	10.30	duration	18.06
	time	7.44	time	16.37
	director/s	7.01	title	14.78
<i>Sports event</i>	type of sport	30.15	type of sport	27.41
	stadium	10.30	name	14.59
	date	12.37	date	13.60
	time	7.44	location	7.01
	winners	6.87	time	5.95
<i>Event</i>	time	17.86	date	18.55
	date	17.31	participants	13.14
	name	10.41	name	12.77
	participants	9.89	duration	8.21
	repetition	6.87	time	7.44

Table C.13: Relevance for EVENT

ARTIFACT				
Category	Attributes (eng)	<i>k</i>	Attributes (it)	<i>k</i>
<i>Product</i>	manufacturer	21.02	function/use	22.52
	price	14.84	name	18.06
	name	14.78	price	14.24
	use	14.33	color	11.13
	warranty	10.30	brand	9.34
<i>Artwork</i>	artist/s	30.91	author	37.77
	creation date	18.43	location	24.57
	style	14.01	style	18.68
	material	10.95	creation date	17.17
	author	7.01	technique	17.17
<i>Building</i>	architect	20.60	number of floors	37.77
	number of floors	17.17	location	32.76
	height	14.90	ara (smq)	20.60
	name	13.13	height	16.37
	architectural style	10.30	date of creation	13.74
<i>Book</i>	author/s	46.71	publisher	61.81
	publisher	44.64	number of pages	54.94
	ISBN	35.63	author	44.34
	year of publication	27.47	title	36.13
	number of pages	24.04	year of publication	30.91
<i>Article of clothing</i>	gender intended for	24.04	size/s	72.11
	color	17.08	brand	35.03
	material	16.42	color	33.39
	style	16.35	fabric	27.47
	fabric	13.74	model	17.17
<i>Object</i>	shape	16.42	color	24.74
	color	13.29	material	22.99
	weight	11.68	shape	22.99
	value	10.30	function/use	21.02
	name	9.85	weight	18.25

Table C.14: Relevance for ARTIFACT

LOCATION				
Category	Attributes (eng)	<i>k</i>	Attributes (it)	<i>k</i>
<i>Tourist Location</i>	name	18.06	attractions	21.93
	attractions	13.74	name	11.49
	geographical position	11.50	services	10.96
	price/s	4.95	population	9.34
	area	4.67	geographical position	9.34
<i>City</i>	population	25.69	population	56.05
	name	21.34	geographical position	23.72
	geographical position	11.50	name	19.70
	region	9.34	region	18.25
	language/s	9.12	climate	14.01
<i>Shop</i>	products sold	13.74	products sold	48.08
	name	13.13	name	31.20
	quality	5.48	timetable	30.15
	owner/s	5.42	number of employees	11.50
	price/s	4.95	location	8.77
<i>Hotel</i>	name	32.84	number of rooms	27.41
	number of rooms	13.74	number of stars	27.41
	rating	10.96	name	24.63
	services	10.30	services	12.45
	number of stars	10.30	category	10.84
<i>Restaurant</i>	type of cuisine	52.08	name	21.34
	name	34.48	timetable	19.19
	chef	17.17	specialty	13.74
	specialty	13.74	type of cuisine	10.30
	price/s	13.60	price/s	9.49
<i>Place</i>	geographical position	42.70	geographical position	29.19
	name	21.34	address	16.82
	continent	13.74	name	11.49
	altitude	10.30	altitude	10.96
	distance from the sea	8.22	continent	8.22

Table C.15: Relevance for LOCATION

## C.2 Entity Search Experiment

### C.2.1 Attribute frequencies for the entity types of the entity search experiment

Category	Attributes	F	f
<i>Politician</i>	surname	53	1
	first name	43	0.81
	role	18	0.40
	location: country	14	0.26
	party	13	0.25
	middle name	5	0.09
	related event	4	0.07
	affiliation	4	0.07
	occupation	4	0.07
	title	2	0.04
	location: city	2	0.04
		N=53	
<i>Manager</i>	surname	38	0.76
	affiliation	30	0.60
	first name	29	0.58
	occupation	20	0.40
	role	13	0.26
	location: country	3	0.26
	location: city	3	0.06
	area of interest/activity	3	0.06
	middle name	2	
	N=50		
<i>Professor</i>	surname	57	0.97
	first name	48	0.81
	affiliation: university	29	0.49
	area of interest/activity	14	0.24
	location: city	13	0.22
	occupation	11	0.18
	affiliation: faculty	7	0.12
	title	6	0.10
	affiliation: institute	4	0.06
	affiliation: department	3	0.05
	N=59		
<i>Sportsperson</i>	surname	48	0.98
	first name	41	0.84
	type of sport	37	0.76
	affiliation: team	18	0.38
	related event	7	0.14
	location: country	4	0.08
	N=49		
<i>Actor/actress</i>	first name	49	0.96
	surname	48	0.94
	movie/s-series	14	0.27
	nationality	6	0.12
	genre	5	0.09
	N=51		
<i>Person</i>	surname	250	0.96
	first name	233	0.90
	occupation	24	0.09
	affiliation	23	0.08
	location: city	15	0.06
	location: country	14	0.05
	area of interest/activity	10	0.04
	middle name	10	0.04
	position/role	6	0.02
	pseudonym	3	0.01
	related event	2	0.01
	famous for	2	0.01
		N=260	

Table C.16: Attribute frequencies in PERSON queries.

<b>Subcategory</b>	<b>Attributes</b>	<b>F</b>	<b>f</b>
<i>Company</i>	name	52	0.74
	type of business	31	0.58
	location: city	12	0.23
	location: country	9	0.17
		N=53	
<i>Association</i>	name	54	1
	location: city	14	0.26
	type of activity	14	0.26
	location: country	5	0.05
		N=54	
<i>University</i>	name	61	1
	location: city	17	0.28
	location: country	8	0.13
	faculties	6	0.09
		N=61	
<i>Government</i>	location: country	48	0.90
	administrative body	11	0.20
	premier	8	0.15
		N=53	
<i>Agency</i>	name	48	0.98
	type/activity	34	0.70
	location: city	13	0.40
	location: country	10	0.16
	location: province	3	0.06
		N=49	
<i>Organization</i> neutral category	name	265	0.95
	type	30	0.11
	activity	27	0.10
	location: city	21	0.08
	location: country	14	0.05
		N=272	

Table C.17: Attribute frequencies in ORGANIZATION queries.



<b>Subcategory</b>	<b>Attributes</b>	<b>F</b>	<b>f</b>
<i>Conference</i>	name	39	0.97
	location: city	16	0.40
	subject	10	0.25
	date: month	5	0.13
	date: year	3	0.07
		N=40	
<i>Meeting</i>	type	27	0.61
	place: city	24	0.54
	name	15	0.34
	date: month	13	0.29
	date: year	10	0.23
	organizers	7	0.16
	subject	5	0.12
		N=44	
<i>Exhibition</i>	location: city	34	0.81
	location: building	13	0.31
	name	13	0.31
	date: year	12	0.28
	subject	9	0.21
	location: country	7	0.17
	date: month	7	0.16
	type	6	0.14
	artist name	6	0.14
		N=42	
<i>Show</i>	type	31	0.70
	name	28	0.63
	location: city	50	
	artist name	12	0.27
	date: year	9	0.20
	location: building	8	0.18
	date: month	4	0.09
		N=44	
<i>Sport event</i>	name	30	0.61
	type of sport	21	0.37
	location: city	21	0.37
	type	12	0.31
	participants	9	0.16
	date: year	9	0.16
	place: country	8	0.14
	date: month	5	0.10
		N=49	
<i>Event</i>	name	159	0.59
	type	125	0.46
	location: city	92	0.33
	artist name	46	0.17
	date: year	24	0.08
	date: month	17	0.06
	location: country	13	0.05
	subjects	10	0.04
	date:day	9	0.03
	location: building	8	0.03
	organizers	5	0.02
		N=271	

Table C.18: Attribute frequencies in EVENT queries.

<b>Subcategory</b>	<b>Attributes</b>	<b>F</b>	<b>f</b>
<i>Product</i>	type	35	0.65
	model name	28	0.52
	brand	25	0.46
	feature	23	0.42
	use	4	0.07
		N=54	
<i>Artwork</i>	title	48	0.92
	creator	22	0.42
	type	15	0.28
	location: museum	14	0.27
	location: city	10	0.19
	style	5	0.09
		N=52	
<i>Building</i>	location: city	44	0.76
	name	42	0.72
	use	16	0.28
	place: country	9	0.16
	type	8	0.14
		N=58	
<i>Book</i>	title	45	0.92
	author	34	0.69
	publisher	6	0.12
	subject	4	0.08
		N=49	
<i>Article of clothing</i>	type	49	0.96
	brand	23	0.45
	material	12	0.24
	features	11	0.21
	sector	9	0.18
	gender intended for	8	0.15
	ways of purchase	5	0.09
		N=51	
<i>Object</i>	type	169	0.63
	name	64	0.24
	model name	49	0.18
	brand	45	0.17
	feature	16	0.06
	creator	14	0.05
			N=269

Table C.19: Attribute frequencies in ARTIFACT queries.

<b>Subcategory</b>	<b>Attributes</b>	<b>F</b>	<b>f</b>
<i>Tourist Location</i>	name	51	0.98
	location: city	14	0.27
	place: country	11	0.21
	sector	7	0.13
	location type	6	0.11
	location: region	6	0.07
		N=52	
<i>City</i>	name	51	0.96
	location: country	25	0.34
	attractions	8	0.15
	administrative role	6	0.11
	location:region	4	0.07
		N=53	
<i>Shop</i>	location: city	35	0.58
	business type	31	0.52
	shop name	31	0.52
	brand	19	0.30
	place: country	5	0.12
		N=60	
<i>Hotel</i>	name	48	0.94
	location: city	42	0.82
	location: country	11	0.21
	location: area	5	0.09
	type	4	0.08
		N=51	
<i>Restaurant</i>	location: city	45	0.86
	name	36	0.69
	type of cuisine	27	0.52
	address	5	0.09
	services	5	0.09
		N=52	
<i>Place</i>	name	255	0.94
	location: city	53	0.19
	location: country	41	0.15
	type	25	0.09
	location: region	13	0.05
		N=271	

Table C.20: Attribute frequencies in LOCATION queries.

## C.2.2 Bayesian relevance measures for low-level entity types

Entity Type ( $e$ )	Attribute type ( $a$ )	$p(e a)$
<i>Politician</i>	party	0.77
	location: country	0.56
	role	0.37
	related event	0.30
	nationality	0.28
	title	0.24
	surname	0.21
	first name	0.20
<i>Manager</i>	occupation	0.55
	affiliation	0.33
	role	0.29
	location: country	0.16
	location: city	0.19
	surname	0.16
	first name	0.15
<i>Professor</i>	location: city	0.57
	title	0.40
	affiliation	0.40
	occupation	0.27
	area of interest/activity	0.21
	surname	0.21
	first name	0.20
<i>Sportsperson</i>	area of interest/activity	0.62
	related event	0.51
	location: country	0.20
	surname	0.21
	first name	0.20
	nationality	0.15
<i>Actor</i>	movies/series	0.79
	role	0.30
	nationality	0.29
	first name	0.24
	surname	0.21

Table C.21: Bayesian Relevance: PERSON

Entity Type ( $e$ )	Attribute type ( $a$ )	$p(e a)$
<i>Company</i>	type	0.40
	activity	0.37
	location: region	0.34
	name	0.24
	location: country	0.21
	location: city	0.17
<i>Association</i>	members	0.42
	location: region	0.33
	type	0.30
	name	0.25
	location: city	0.20
<i>University</i>	faculties	0.60
	name	0.24
	location: city	0.21
	location: country	0.17
<i>Government</i>	administrative body	0.75
	premier	0.69
	location: country	0.23
	location: city	0.19
<i>Agency</i>	activity	0.43
	location: country	0.25
	location: city	0.20

Table C.22: Bayesian Relevance: ORGANIZATION

<b>Entity Type (<math>e</math>)</b>	<b>Attribute type (<math>a</math>)</b>	<b><math>p(e a)</math></b>
<i>Conference</i>	subject	0.39
	name	0.39
	location: country	0.22
	organizers	0.21
	date: month	0.18
<i>Meeting</i>	organizers	0.56
	date: month	0.42
	type	0.39
	name	0.25
	date: year	0.28
	location: city	0.22
<i>Exhibition</i>	location: country	0.57
	location: building	0.55
	date: day	0.43
	subject	0.36
	location: city	0.34
	date: year	0.33
	artist name	0.31
	date: month	0.21
<i>Show</i>	artist name	0.59
	type	0.44
	location: building	0.35
	date: day	0.32
	name	0.28
	location: city	0.22
<i>Sport event</i>	type of sport	0.81
	location: country	0.35
	date: day	0.23
	name	0.22
	date: year	0.18

Table C.23: Bayesian Relevance: EVENT

<b>Entity Type (<math>e</math>)</b>	<b>Attribute type (<math>a</math>)</b>	<b><math>p(e a)</math></b>
<i>Product</i>	model	0.89
	brand	0.72
	feature	0.71
	type	0.45
	use	0.20
<i>Artwork</i>	location: building	0.82
	style	0.62
	nationality	0.55
	creator	0.37
	title/name	0.35
	type	0.20
<i>Building</i>	location: city	0.71
	use	0.70
	location: country	0.55
	name	0.30
<i>Book</i>	publisher	0.68
	subject	0.60
	creator	0.56
	title/name	0.33
<i>Article of Clothing</i>	sector	0.42
	gender intended for	0.39
	material	0.38
	brand	0.20

Table C.24: Bayesian Relevance: ARTIFACT

<b>Entity Type (<math>e</math>)</b>	<b>Attribute type (<math>a</math>)</b>	<b><math>p(e a)</math></b>
<i>Tourist location</i>	location name	0.74
	location type	0.28
	organization name	0.18
<i>City</i>	administrative role	0.68
	building name	0.68
	state name	0.48
	municipality	0.48
	country name	0.46
	city name	0.30
<i>Shop</i>	shop name	0.91
	product type	0.90
	brand	0.85
	shop type	0.79
	address:street	0.33
<i>Hotel</i>	hotel name	0.93
	hotel type	0.61
	number of stars	0.48
	price range	0.42
<i>Restaurant</i>	restaurant name	0.92
	type of cuisine	0.90
	restaurant type	0.61
	services	0.47
	location: neighbourhood	0.43

Table C.25: Bayesian Relevance: LOCATION

### C.2.3 Position Distribution of Attribute Types

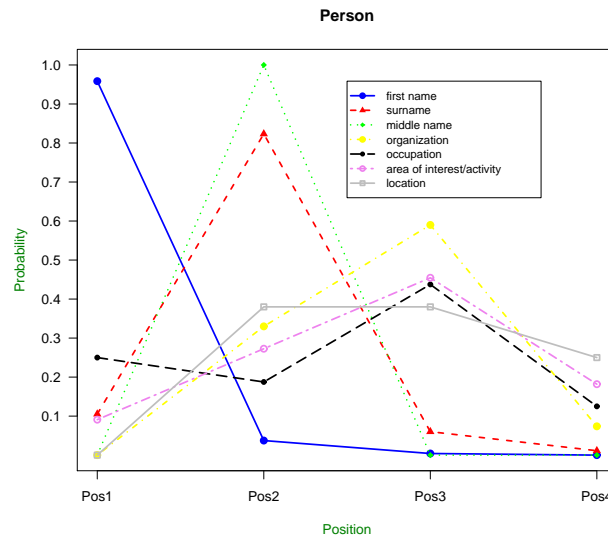


Figure C.1: Probability distribution of attribute types for the first four positions in queries about Person.

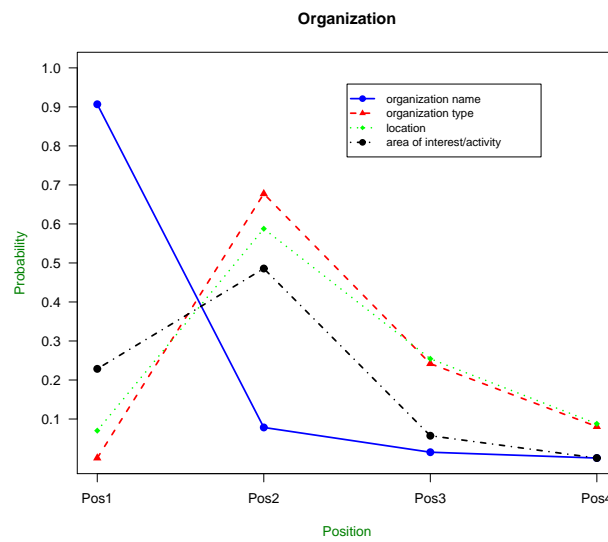


Figure C.2: Probability distribution of attribute types for the first four positions in queries about Organization.

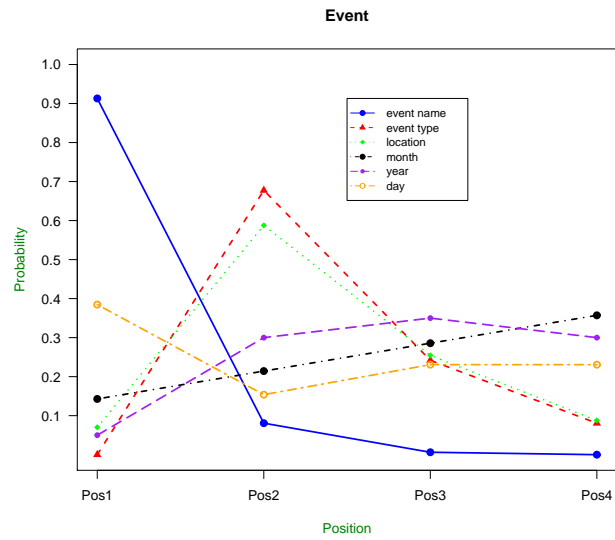


Figure C.3: Probability distribution of attribute types for the first four positions in queries about Event.

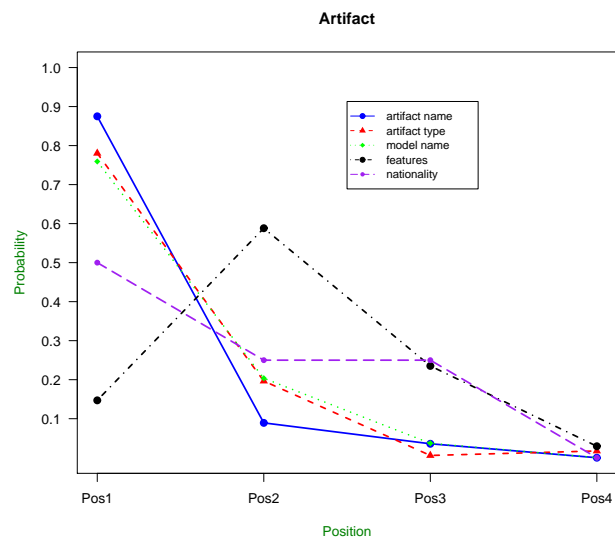


Figure C.4: Probability distribution of attribute types for the first four positions in queries about Artifact.



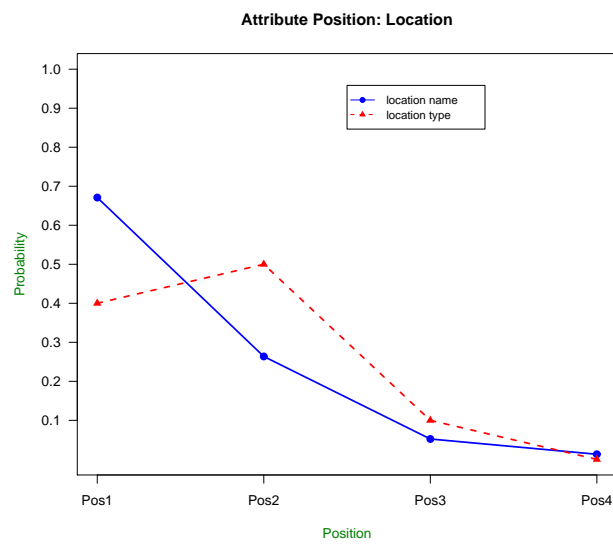


Figure C.5: Probability distribution of attribute types for the first four positions in queries about Location.

## Appendix D

# Mutability and Causality Ratings: Stimuli and Measures

### D.1 Entity profiles used to collect mutability and causality ratings

The following profiles were used to collect mutability and causality ratings.

#### 1) PERSON DESCRIPTIONS

**NAME: Madison Smith**

AGE: 45

HOBBIES: tennis

OCCUPATION: reporter

PHONE: 202.287.3305

HEIGHT: 5'8"

**NAME: Michael Abrams**

BIRTH DATE: July 15, 1969

EXPERIENCES AND QUALIFICATIONS: master in business administration, Stanford Advanced  
Project Management certificate

WEIGHT: 170 pounds

CITY OF RESIDENCE: London

EMAIL: abrams@gmail.com

**NAME: Benjamin Green**

RELIGION: catholic

BIRTH PLACE: Los Angeles

EYE COLOR: blue

EDUCATION: master degree in law

ADDRESS: 4300 Hudson Ave.

**NAME: Nathan McConnell**

NATIONALITY: Ireland

MARITAL STATUS: single  
SOCIAL SECURITY NUMBER: 595-12-5274  
AFFILIATION: Porter Airlines  
HAIR COLOR: blonde

**NAME: Nicholas Patton**  
CHARACTER: sociable  
MOTHER'S NAME: Julia Anderson  
ROLE: editor in chief  
COUNTRY OF RESIDENCE: Germany  
GENDER: male

## 2) ORGANIZATION DESCRIPTIONS

International Organization

**NAME: Asian Commission for Environmental Cooperation**  
MEMBER STATES: China, India, Japan, Australia, Nepal, Mongolia  
HEADQUARTERS (location): Bangkok  
LEGAL STATUS: non governmental  
DATE OF FOUNDATION: 1988  
WEB SITE URL: <http://www.acec.com>

Company

**NAME: Cyber**  
COUNTRY: USA  
BUSINESS TYPE: computer hardware and software  
CEO: John Anderson  
OWNER: Michael Thomson  
ANNUAL TURNOVER: 2 million dollars

Association

**NAME: APAP**  
CITY: Los Angeles  
MAIN OBJECTIVE: prevention and health promotion  
PRESIDENT: Carol Walton  
ASSOCIATES: Robert Burton, Alexander Luan  
EMAIL: [apap@gmail.com](mailto:apap@gmail.com)

Agency

NAME: Job Finder  
PRESIDENT: John Langton  
ADDRESS: 1345 Polaris Ave  
TYPE: employment agency  
NUMBER OF EMPLOYEES: 50  
PHONE NUMBER: 345 678956

University

**NAME: Cajal University**  
CITY: Madrid  
FACULTIES: History, International Business, Philosophy  
DEPARTMENTS: Departments of Business Administration and Economics . . . Department of Human Sciences  
NUMBER OF STUDENTS: 11000  
COURSES: Introduction to American Studies . . . Modern Philosophy

### 3) EVENT DESCRIPTIONS

Conference

**NAME: ACS**

LOCATION: 14750 Conference Center Drive

DATES: October 25-29

YEAR: 2009

ORGANIZERS: Lisa Zhang, Mike Dean

SPONSORS: BBN Technology, ManTech

Show

**NAME: Rainbow**

CITY: New York

TYPE: musical

DATE: September 4-6

LEAD ACTOR: David Alvarez

DIRECTOR: Stephen Daldry

Sport event

**NAME: Beach World Cup**

COUNTRY: Dubai

YEAR: 2009

TIME: 13 p.m.

TYPE OF SPORT: Soccer

PARTICIPANTS: Mexico, Italy; Germany, France

Meeting

**NAME: AIB (Academy of International Business)**

TIME: 9 a.m. - 4 p.m.

DATE: October 25

YEAR: 2009

LOCATION: College of Business Administration, San Diego State University

TOPIC: business strategies during the depression

Exhibition

**NAME: The Unusual Object**

SUBJECT: surrealism and the power of the imagination to transform the everyday

LOCATION: The Museum of Modern Art, 11 West 53 Street

DATES: June 24, 2009-January 4, 2010

PRICE: \$20

MAIN ARTISTS: Salvador Dalí and Meret Oppenheim

### 4) ARTIFACT DESCRIPTIONS

Product

**NAME: Wing**

PRICE: \$449

MODEL: gd900

COLOR: black

SIZE: 13"

MANUFACTURER: Sony

Book

**TITLE: Mind Shadow**  
AUTHOR: Richard Cabot  
PUBLISHER: Dell Publishing  
NUMBER OF PAGES: 192  
ISBN: 9780440204886  
EDITION: 1989

Artwork

**TITLE: Still Life**  
AUTHOR: Patricia Waddell  
DISPLAY LOCATION: San Francisco Museum of Modern Art  
STYLE: contemporary art  
CREATION DATE: 1990  
MATERIAL: oil on canvas

Building

**NAME: Green Hall**  
ADDRESS: 3131 McClintock Ave  
HEIGHT: 50m  
USE: residence hall  
COLOR: white  
ARCHITECT: Jackie Craven

Software

**NAME: Photo Power**  
TYPE: freeware  
VERSION: 3.3  
FEATURES/FUNCTIONS: photo editing, screen capturing, raw converter  
MANUFACTURER: MOOII TECH  
CREATOR: Paul Griffin

#### 5) LOCATION DESCRIPTIONS

City

**NAME: Eastport**  
POPULATION: 1000  
GEO COORDINATES: 44°54'49"N 67°0'14"W  
STATE/REGION: Washington County, Maine  
AREA: 13 sq mi  
MAIN LANGUAGE: English

Shop

**NAME: Calibre**  
ADDRESS (street): 139 Elizabeth St.  
SHOP TYPE: retail store  
PRODUCTS SOLD: sport clothes  
OPENING HOURS: 9 a.m.- 6 p.m.  
AREA: 150 mq

Restaurant

**NAME: Khaosan**

COUSINE TYPE: Thai Restaurant  
 CITY: London  
 PRICE RANGE: \$20-50  
 CHEF: Win Liaowarin  
 RATING: medium

Tourist location

**NAME: Fun Spot**  
 ADDRESS: 5551 Del Verde Way, Orlando  
 TYPE: amusement park  
 PRICE: \$35  
 HOURS: 10:00 - midnight  
 ATTRACTION/S: Ferris wheel

Hotel

**NAME: Plaza**  
 ADDRESS (street): 1345 Richmond Street  
 NUMBER OF ROOMS: 300  
 NUMBER OF STARS: 4  
 SERVICES: Concierge, High-speed Internet access  
 OWNER: Terry Tailor

## D.2 Mutability and Causality Ratings

Entity Type	Attributes	Mutability Ratings		Causality Ratings	
		Mean	SD	Mean	SD
Person 1	age	4.93	3.17	5.5	3.7
	hobbies	4.81	2.04	3.7	2.54
	occupation	4.98	2.17	3.8	2.31
	phone number	1.81	2.07	2.75	2.32
	height	4.15	2.29	6.68	2.08
Person 2	birth date	6.44	2.78	8.68	0.70
	qualifications	5.75	2.17	4.81	2.78
	weight	2.31	1.40	2.56	2.44
	city of residence	3.43	2.22	3.12	2.27
Person 3	email	1.25	0.57	3.06	2.88
	religion	5.68	2.70	5.25	2.46
	birth place	5.12	2.60	8.62	0.80
	eye color	4.12	2.39	5.18	3.20
	education	6.12	1.92	3.37	2.27
Person 4	address	2.0	0.96	2.56	2.12
	nationality	7.06	2.20	6.81	2.83
	marital status	4.25	2.88	2.06	1.91
	social security number	4.31	3.64	8.18	1.04
	affiliation	3.0	1.89	3.0	2.06
Person 5	hair color	2.31	1.53	2.43	2.52
	character	6.43	2.06	4.25	2.93
	mother's name	4.93	3.08	5.81	3.35
	role	4.31	2.15	3.12	2.47
	country of residence	4.12	2.33	3.18	2.10
Int. Organization	gender	8.06	1.94	6.25	2.56
	member states	6.06	2.04	6.93	2.69
	headquarters	3.93	2.37	3.93	2.29
	legal status	6.06	1.98	4.75	2.51

	date of foundation	4.93	2.64	8.25	1.69
	web site url	2.25	2.11	4.25	2.86
Company	Country	5.06	2.79	4.75	2.67
	business type	8.0	0.89	6.06	2.23
	ceo	5.19	2.34	2.75	2.01
	owner	4.81	2.13	2.87	2.06
	annual turnover	4.50	2.22	3.18	1.93
Association	city	4.37	2.18	5.06	2.23
	main objective	8.12	1.31	5.75	2.35
	president	4.25	2.11	2.12	1.85
	associates	4.75	1.69	3.18	2.63
	email	1.37	0.61	5.75	1.98
Agency	president	3.75	2.20	2.12	1.85
	address	2.93	1.56	3.18	2.63
	type	8.00	1.15	5.75	1.98
	number of employees	3.56	1.96	2.37	1.82
	phone number	1.31	0.60	3.37	2.47
University	city	6.62	1.58	7.5	1.71
	faculties	6.31	2.24	6.18	2.31
	departments	6.25	1.87	6.75	2.65
	number of students	2.68	1.81	2.75	2.62
	courses	4.87	2.33	6.93	2.14
Conference	location	3.87	2.27	3.25	2.20
	dates	2.50	1.82	3.5	1.93
	year	5.18	2.16	4.93	2.32
	organizers	5.75	1.57	6.37	2.39
	sponsors	3.62	2.06	5.43	2.55
Show	city	4.43	2.18	3.31	2.18
	type	7.81	1.47	6.25	2.23
	date	2.68	1.70	3.37	2.02
	lead actor	6.25	2.48	3.93	1.94
	director	6.0	2.09	3.87	2.36
Sport Event	country	5.5	2.03	4.06	2.56
	year	5.12	2.82	6.68	2.05
	time	2.62	2.12	2.50	1.09
	type of sport	8.50	1.03	7.00	2.44
	participants	4.93	2.64	5.56	2.82
Meeting	time	2.50	1.71	3.18	2.63
	date	2.62	1.66	3.56	2.50
	year	5.06	2.26	6.00	2.55
	location	2.81	1.90	3.56	2.50
	topic	7.37	1.85	4.68	2.30
Exhibition	subject	7.81	0.98	5.37	2.36
	location	4.37	1.85	3.00	1.54
	dates	2.68	2.08	3.37	2.24
	price	3.06	1.80	2.37	1.20
	main artists	7.75	1.0	6.06	2.51
Product	price	3.37	2.50	1.93	0.92
	model	6.31	2.02	4.37	2.98
	color	3.31	2.05	2.81	2.37
	size	6.12	1.54	4.43	2.94
	manufacturer	4.87	2.15	4.68	2.82
Book	author	8.50	0.81	8.43	0.96
	publisher	4.25	2.79	6.12	2.70
	number of pages	4.25	2.67	4.81	3.20
	ISBN	4.56	3.59	5.62	2.77
	edition	3.93	2.51	5.06	3.45
Artwork	author	8.31	1.01	8.43	0.81
	display location	3.25	2.40	2.43	1.78

	style	8.00	0.96	7.37	1.82
	creation date	4.93	2.88	8.00	1.82
	material	7.81	1.32	7.75	2.08
Building	address	4.93	2.67	5.12	2.91
	height	5.5	1.93	4.43	2.65
	use	4.93	2.93	4.00	1.96
	color	2.75	1.77	2.62	2.30
	architect	6.87	2.06	7.37	2.33
Software	type	5.87	2.15	4.93	2.93
	version	4.81	1.90	3.93	2.83
	functions	8.43	0.89	6.93	2.29
	manufacturer	3.87	1.82	5.56	2.03
	creator	7.0	1.63	7.81	1.97
City	population	4.81	2.71	2.06	1.18
	geo coordinates	7.06	1.91	8.25	1.80
	state	8.31	1.35	7.81	2.19
	area	4.56	2.25	5.81	2.50
	main language	6.93	1.98	7.56	2.52
Shop	address	3.56	1.82	3.06	2.46
	shop type	7.25	2.26	4.93	1.28
	product sold	7.68	1.4	4.81	2.19
	opening hours	2.68	1.81	2.12	1.31
	area	4.81	2.07	2.81	2.25
Restaurant	cuisine type	7.62	1.58	5.87	2.39
	city	5.31	1.85	4.93	2.56
	price range	4.5	1.89	3.00	1.26
	chef	5.31	2.67	3.43	2.47
	rating	5.25	2.48	2.93	1.65
	address	5.37	2.06	4.68	2.07
Tourist Location	type	8.0	1.82	5.5	2.36
	price	3.43	2.12	2.62	1.62
	hours	2.68	1.95	3.25	2.74
	main attractions	6.37	2.24	4.62	2.30
Hotel	address	3.93	2.04	4.87	2.52
	number of rooms	4.75	2.56	4.25	2.26
	number of stars	5.31	2.54	3.00	1.71
	services	5.43	2.33	6.37	2.52
	owner	4.31	2.24	3.31	2.54

Table D.1: Mean (and SD= standard deviation) Mutability and Causality Ratings

## D.3 Entity Profiles used in Experiment 2

ORIGINAL	CONTINUER 1	CONTINUER 2
NAME: Carol Green	Carol Green	Carol Green
Address: 506 South Grand Ave	701 Pennsylvania Ave	506 South Grand Ave
Weight: 126lb	126lb	136lb
NAME: Robert Smith	Robert Smith	Robert Smith
Hair color: brown	red	brown
City of residence: New York	New York	Chicago
NAME: Stephen Young	Stephen Young	Stephen Young
Phone number: 312-263-1737	847-125-1007	312-263-1737
Occupation: professor	professor	financial counselor
NAME: Nathan McConnell	Nathan McConnell	Nathan McConnell
Weight: 160 lb	150 lb	160 lb
Religion: Catholic	Catholic	Buddhism
NAME: Rachel James	Rachel James	Rachel James



<i>Marital status:</i> single	married	single
<i>Birth date:</i> 29 September 1966	29 September 1966	31 August 1972
NAME: Bob James	Bob James	Bob James
<i>City of residence:</i> New York	Toronto	New York
<i>Affiliation:</i> IMSI	IMSI	Visio
NAME: Mary Scott	Mary Scott	Mary Scott
<i>Email:</i> mary@gmail.com	chubby@yahoo.com	mary@gmail.com
<i>Hobbies:</i> gardening	gardening	stamp collecting
NAME: Michael Abrams	Michael Abrams	Michael Abrams
<i>Role:</i> Business Administrator	Human resource manager	Business Administrator
<i>Age:</i> 46	46	41
NAME: Sarah Randolph	Sarah Randolph	Sarah Randolph
<i>Country of residence:</i> Germany	Ireland	Germany
<i>Height:</i> 5'6"	5'6"	5'3"
NAME: Carl Larson	Carl Larson	Carl Larson
<i>Hobbies:</i> tennis	kayaking	tennis
<i>Character:</i> sociable	sociable	antisocial
NAME: Anna Jones	Anna Jones	Anna Jones
<i>Qualifications:</i> public adm. certificate	business adm. certificate	public adm. certificate
<i>Eye color:</i> brown	brown	green
NAME: Virginia Tylor	Virginia Tylor	Virginia Tylor
<i>Occupation:</i> hairdresser	shop assistant	hairdresser
<i>Birth date:</i> 1 January 1968	1 January 1968	25 May 1971
NAME: Madison William	Madison William	Madison William
<i>Gender:</i> male	female	male
<i>Mother's name:</i> Alyssa Thomson	Alyssa Thomson	Emma Paxton
NAME: David Smith	David Smith	David Smith
<i>Religion:</i> Catholic	Jewish	Catholic
<i>Social security num:</i> 431-45-9876	431-45-9876	123-46-6789
NAME: Alexandra Brown	Alexandra Brown	Alexandra Brown
<i>Height:</i> 5'40"	5' 50"	5'40"
<i>Birth place:</i> Miami	Miami	Santa Barbara

Table D.2: Person Profiles used in the experiment2

ORIGINAL	CONTINUER 1	CONTINUER 2
NAME: Cyber	Cyber	Cyber
<i>Owner:</i> Michael Thomson	Robert Lewis	Michael Thomson
<i>Ceo:</i> John Anderson	John Anderson	Anthony Moore
NAME: Horizon	Horizon	Horizon
<i>Address:</i> 1345 Boston Ave	1915 Polaris Ave	1345 Boston Ave
<i>President:</i> Carol Walton	Carol Walton	Emma Johnson
NAME: Biotech	Biotech	Biotech
<i>Owner:</i> Addison Foster	Robert Lewis	Addison Foster
<i>Country:</i> United Kingdom	United Kingdom	Germany
NAME: Cajal University	Cajal University	Cajal University
<i>N. students:</i> 10000	13000	10000
<i>Departments:</i> Economics Anthropology, Chemistry	Economics, Anthropology, Chemistry	Physics Sociology, Philosophy
NAME: Pauling University	Pauling University	Pauling University
<i>N. students:</i> 15000	13000	15000
<i>City:</i> Berlin	Berlin	Frankfurt
NAME: Friendship Charity Ass.	Friendship Charity Ass.	Friendship Charity Ass.
<i>President:</i> Meredith Baxter	Carol Lee	Meredith Baxter
<i>Email:</i> info@fshipcharity.org	info@fshipcharity.org	contact@fca.org
NAME: Youth Football Ass.	Youth Football Ass.	Youth Football Ass.
<i>Email:</i> info@yfa.org	arniex@virgin.net	info@yfa.org
<i>City:</i> San Francisco	San Francisco	San Diego

NAME: SCA <i>President:</i> Carol Walton <i>Associates:</i> Brett Kleist, A. Tang, M. Esseman	SCA Larry Christiansen Brett Kleist, A. Tang, M. Esseman	SCA Carol Walton Stuart Finney, M. Phelps, A. Wang
NAME: Third eye <i>Phone number:</i> 212-509-7200 <i>Type:</i> advertising agency	Third eye 212-777-7534 advertising agency	Third eye 212-509-7200 employment agency
NAME: For the Right to Food <i>Headquarters:</i> Paris <i>Legal status:</i> non governmental	For the Right to Food Amsterdam non governmental	For the Right to Food Paris intergovernmental
NAME: COA <i>City:</i> Columbus <i>Main objective:</i> obesity prevention	COA Pittsburgh obesity prevention	COA Columbus promote and coordinate chess activities
NAME: Commission on Climate Change <i>Headquarters:</i> Shanghai <i>Member states:</i> China, India, Japan	Commission on Climate Change Tokyo China, India, Japan	Commission on Climate Change Shanghai Nepal, Australia, Vietnam
NAME: IAOP <i>Objective:</i> ent. older people <i>Associates:</i> Robert O'Neill, Mary Lynch, Linda King	IAOP sup. autistic peo. Robert O'Neill, Mary Lynch, Linda King	IAOP ent. older peo. Alan Scott, Barbara Hogan, Marc Reid
NAME: Fermi University <i>Departments:</i> Biology, Economics, Engineering <i>City:</i> Rome	Fermi University Anthropology, Chemistry, Physics Rome	Fermi University Art and Sciences, Economics, Engineering Venice
NAME: IFA <i>Member states:</i> Texas, Maine, Georgia <i>Date of foundation:</i> 25-10-1980	IFA Florida, Virginia, Ohio 25-10-1980	IFA Texas, Arkansas, Georgia 2-11-1995

Table D.3: Organization Profiles used in the experiment2

ORIGINAL	CONTINUER 1	CONTINUER 2
NAME: The Unusual Object <i>Location:</i> Museum of Modern Art <i>Price:</i> \$20	The Unusual Object Axia Modern Art \$20	The Unusual Object Museum of Modern Art \$15
NAME: Asian Semantic Web Conf. <i>Location:</i> Pierre Baudis Center <i>Dates:</i> November 25-28	Asian SW Conference Sir Alexander Fleming build. November 25-28	Asian SW Conference Pierre Baudis Center September 12-17
NAME: Breeders' Cup <i>Time:</i> 9 a.m - 4 p.m. <i>Country:</i> Canada	NAME: Breeders' Cup 11 a.m - 6 p.m. Canada	NAME: Breeders' Cup 9 a.m - 4 p.m. USA
NAME: Business Forum <i>Time:</i> 9 a.m - 13 p.m. <i>Year:</i> 2009	Business Forum 14p.m - 17 p.m. 2009	Business Forum 9 a.m - 13 p.m. 2010
NAME: Currie Cup <i>Time:</i> 9 a.m - 18 p.m. <i>Type of sport:</i> rugby	Currie Cup 15 a.m - 21 p.m. rugby	Currie Cup 9 a.m - 18 p.m. bowling
NAME: Mary Poppins <i>City:</i> New York <i>Date:</i> October 25-29	Mary Poppins Amsterdam October 25-29	Mary Poppins New York December 25-29
NAME: The First Dream <i>City:</i> Chicago <i>Lead actor:</i> Steven Robman	The First Dream Washington Steven Robman	The First Dream Chicago Henry Condell
NAME: NanoThech	NanoThech	NanoThech

<i>Dates:</i> June 23-26	July 1-4	June 23-26
<i>Sponsors:</i> NSTI, CTSI, Platinum	NSTI, CTSI, Platinum	NanoInk, Merck, BASF
NAME: Infringe the Obvious	Infringe the Obvious	Infringe the Obvious
<i>Dates:</i> June 24, 2009 - January 4, 2010	October 12, 2009 - February 2, 2010	June 24, 2009 - January 4, 2010
<i>Main artists:</i> C. Lim, B. Puah	C. Lim, B. Puah	J. Hiah, F. West
NAME: The Misanthrope	The Misanthrope	The Misanthrope
<i>Lead actor:</i> Melanie Klein	Kelly Price	Melanie Klein
<i>Director:</i> Thea Sharrock	Thea Sharrock	Tom Morris
NAME: ABS	ABS	ABS
<i>Topic:</i> business strategy	safety	business strategy
<i>Year:</i> 2009	2009	2010
NAME: Rip Curl Pro	Rip Curl Pro	Rip Curl Pro
<i>Country:</i> Australia	New Zealand	Australia
<i>Type of sport:</i> surf	surf	horse race
NAME: GECCO	GECCO	GECCO
<i>Year:</i> 2008	2009	2008
<i>Sponsors:</i> Toyota, Philips	NSTI, CTSI, Platinum	Nvidia, Icosystem Corporation
NAME: Beach Soccer Festival	Beach Soccer Festival	Beach Soccer Festival
<i>Participants:</i> Italy, Japan, Senegal	Germany , Australia, China	Italy, Japan, Senegal
<i>Year:</i> 2007	2007	2008
NAME: Anaconda	Anaconda	Anaconda
<i>Year:</i> 2009	2010	2009
<i>Type of sport:</i> surf	surf	triathlon

Table D.4: Event Profiles used in the experiment2

ORIGINAL	CONTINUER 1	CONTINUER 2
NAME: Viparis	Viparis	Viparis
<i>Use:</i> convention center	shopping center	convention center
<i>Color:</i> grey	grey	white
NAME: Wing	NAME: Wing	NAME: Wing
<i>Price:</i> \$200	\$250	\$200
<i>Model:</i> gd900	gd900	gd1100
NAME: Margot Guest House	Margot Guest House	Margot Guest House
<i>Color:</i> yellow	green	yellow
<i>Address:</i> 1678 Lexington Ave	1678 Lexington Ave	317 West 14th Street
NAME: Remembrance	Remembrance	Remembrance
<i>Display location:</i> High Museum of Art	Museum of Fine Arts	High Museum of Art
<i>Style:</i> expressionism	expressionism	cubism
NAME: Embrace	Embrace	Embrace
<i>Display location:</i> National Museum of Modern Art	Art Institute	National Museum of Modern Art
<i>Artist:</i> Michelle Ward	Michelle Ward	Megan Faye
NAME: XEL	XEL	XEL
<i>Model:</i> KDL-46X3500	RG-250	KDL-46X3500
<i>Size:</i> 32"	32"	46"
NAME: Left Back	Left Back	Left Back
<i>Number of pages:</i> 250	270	250
<i>Edition:</i> 2th	2th	4th
NAME: Amulet	Amulet	Amulet
<i>Version:</i> 1.1	1.5	1.1
<i>Function:</i> photo editing	photo editing	file sharing
NAME: Codec	Codec	Codec
<i>Version:</i> 2.1.2	1.1.4	2.1.2

<i>Creator:</i> Andrew Johnson	Andrew Johnson	Ian Darragh
<i>NAME:</i> Akismet	Akismet	Akismet
<i>Type:</i> freeware	commercial software	freeware
<i>Manufacturer:</i> Zone Labs	Zone Labs	TackTech
<i>NAME:</i> The Soul’s Darkness	The Soul’s Darkness	The Soul’s Darkness
<i>Edition:</i> 1th	2th	1th
<i>Publisher:</i> Picador	Picador	Henry Holt and Co.
<i>NAME:</i> Mid Town Tower	Mid Town Tower	Mid Town Tower
<i>Address:</i> 1678 Zola Ave	1915 Sherman Way	1678 Zola Ave
<i>Architect:</i> David Fisher	David Fisher	Helmut Jahn
<i>NAME:</i> Last Moon	Last Moon	Last Moon
<i>Publisher:</i> Little, Brown & Company	Samhain Publishing	Little, Brown & Company
<i>ISBN:</i> 978-0316166317	978-0316166317	978-1599982595
<i>NAME:</i> Bridgit	Bridit	Bridit
<i>Functions:</i> audio conferencing	video conferencing	audio conferencing
<i>Creator:</i> Paul Barlow	Paul Barlow	Nicholas Chapman
<i>NAME:</i> Santa Cruz	Santa Cruz	Santa Cruz
<i>Author:</i> Paul Bloomer	Michelle Ward	Paul Bloomer
<i>Creation date:</i> April 1980	April 1980	August 1992

Table D.5: Artifact Profiles used in the experiment2

<b>ORIGINAL</b>	<b>CONTINUER 1</b>	<b>CONTINUER 2</b>
<i>NAME:</i> Khnumhotep’s Tomb	Khnumhotep’s Tomb	Khnumhotep’s Tomb
<i>Price:</i> \$20	\$25	\$20
<i>Hours:</i> 10:00 am - 5:00 pm	10:00 am - 5:00 pm	2:00 pm - 5:00 pm
<i>NAME:</i> Petal	Petal	Petal
<i>Area:</i> 80smq	130sqm	80smq
<i>Address:</i> 1661 York Ave	1661 York Ave	1345 Madison Ave
<i>NAME:</i> Old Town	Old Town	Old Town
<i>Price:</i> \$30	\$20	\$30
<i>Address:</i> 834 Surf Ave	834 Surf Ave	11131 Malibu Dr
<i>NAME:</i> Antigone	Antigone	Antigone
<i>Opening hours:</i> 10:00 am - 9:00 pm	11:00 am - 7:00 pm	10:00 am - 9:00 pm
<i>Shop type:</i> outlet	outlet	department store
<i>NAME:</i> Grafton	Grafton	Grafton
<i>Population:</i> 56,257	35,000	56,257
<i>Geo coordinates:</i> 29°41’S 152°56’E	29°41’S 152°56’E	33°51’S 151°12’E
<i>NAME:</i> Grace Hotel	Grace Hotel	Grace Hotel
<i>Number of stars:</i> 3	5	3
<i>Owner:</i> Cherie Ditcham	Cherie Ditcham	Alex Scott
7. <i>NAME:</i> The Pavilion	The Pavilion	The Pavilion
<i>Price range:</i> \$40-70	\$80-120	\$40-70
<i>Chef:</i> Alex Di Maggio	Alex Di Maggio	Antonio Tettamanzi
<i>NAME:</i> Carmelita	Carmelita	Carmelita
<i>Rating:</i> 4.5	3.8	4.5
<i>Cuisine type:</i> Mexican	Mexican	Seafood
<i>NAME:</i> Capital Hotel	Capital Hotel	Capital Hotel
<i>Number of stars:</i> 5	4	5
<i>Services:</i> Wi-Fi access in public areas	Wi-Fi access in public areas	In room Wi-Fi
<i>NAME:</i> Heritage Park	Heritage Park	Heritage Park
<i>Address:</i> 861 SE Main Street	2470 Heritage Park Row	861 SE Main Street
<i>Main attraction:</i> Neptune Fountain	Neptune Fountain	Steam Locomotive
<i>NAME:</i> Euro Queen Hotel	Euro Queen Hotel	Euro Queen Hotel
<i>Number of rooms:</i> 196	150	196
<i>Address:</i> 122 Church Road	122 Church Road	110 Peckham Road
<i>NAME:</i> Griffin House	Griffin House	Griffin House
<i>Number of rooms:</i> 25	50	25

<i>Services:</i> 24-hour front desk	24-hour front desk	daytime front desk
NAME: Sitar	Sitar	Sitar
<i>Cuisine type:</i> Indian	Japanese	Indian
<i>City:</i> Boston	Boston	Milwaukee
NAME: Fantasy	Fantasy	Fantasy
<i>Address:</i> 1101 Van Ness Ave	122 Church Road	1101 Van Ness Ave
<i>Services:</i> free parking	free parking	valet parking
NAME: Queensburg	Queensburg	Queensburg
<i>Area:</i> 157 Km2	165 Km2	157 Km2
<i>Main language:</i> English	English	French

Table D.6: Location Profiles used in the experiment2

## D.4 Response Distribution in Experiment 2

### D.4.1 Causal Continuer Model Fit

Graphical representations of the Causal Continuer Model fitting for the 15 trials of the five categories used in experiment 2.

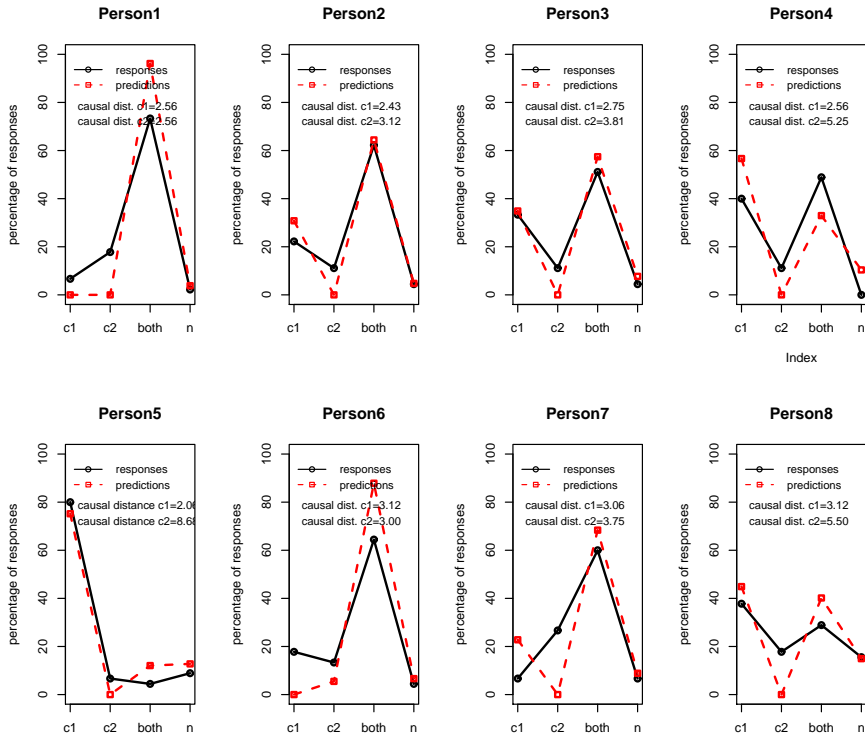


Figure D.1: Person tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

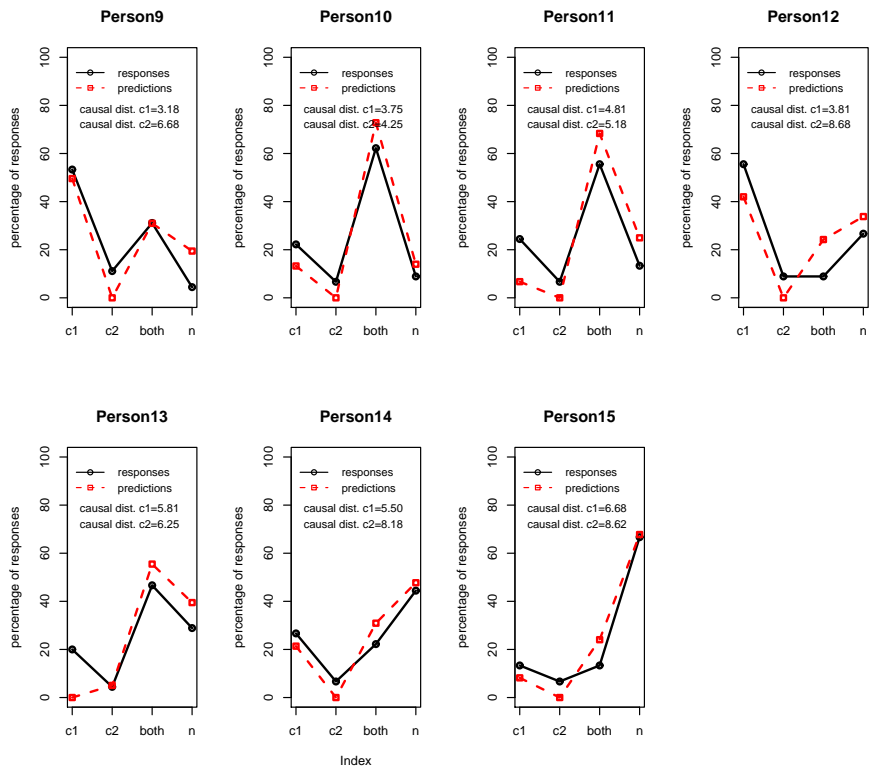


Figure D.2: Person tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

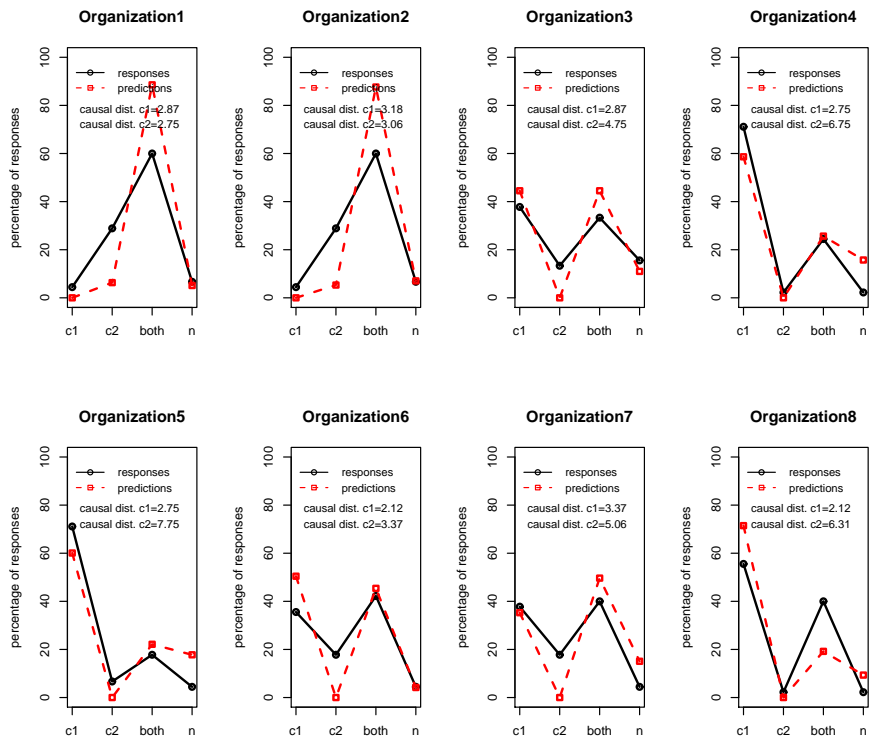


Figure D.3: Organization tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

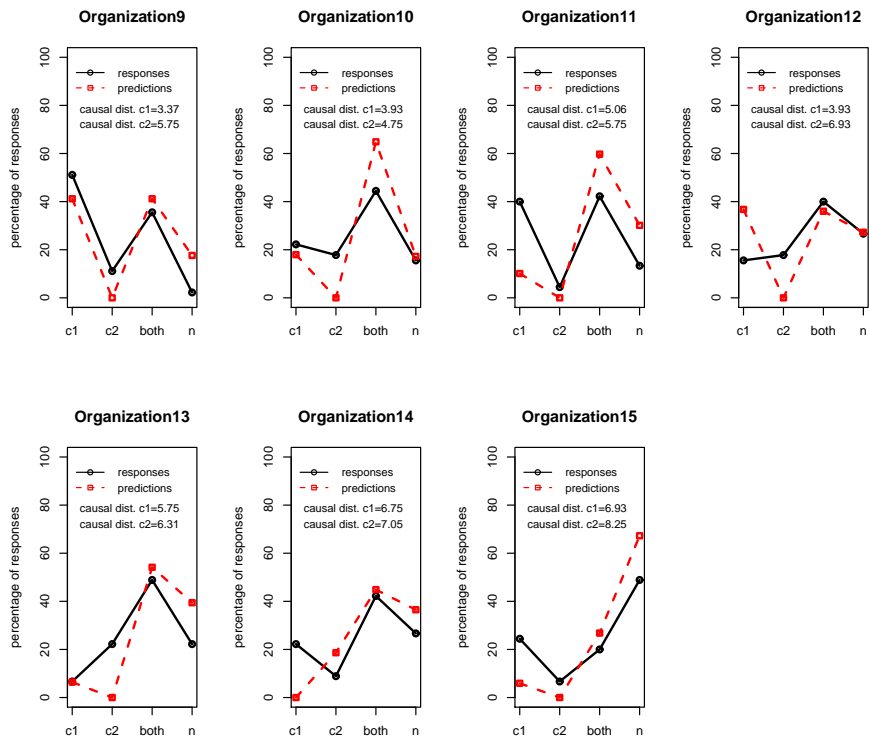


Figure D.4: Organization tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.



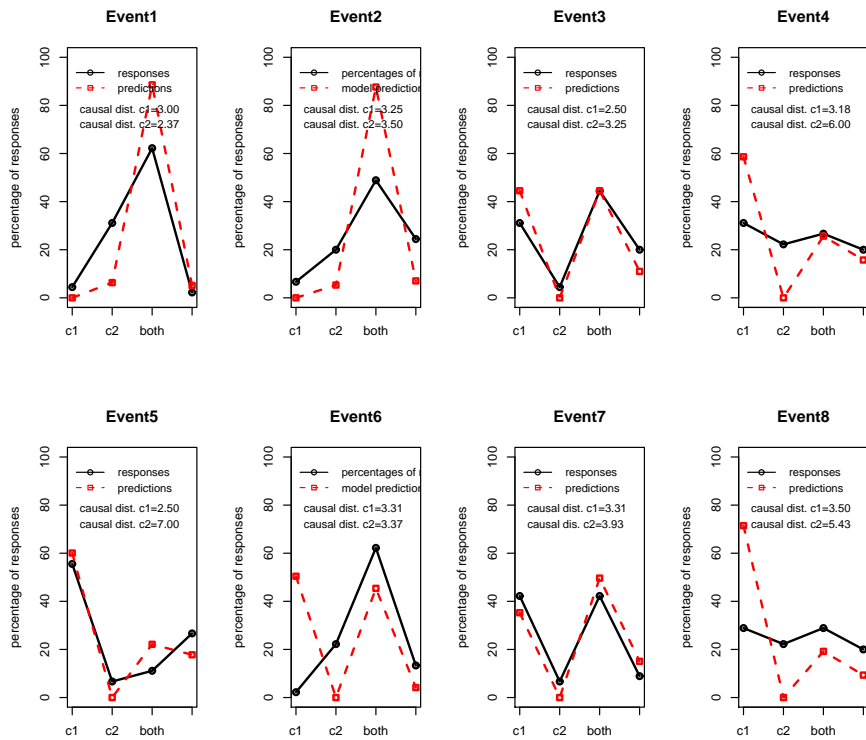


Figure D.5: Event tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

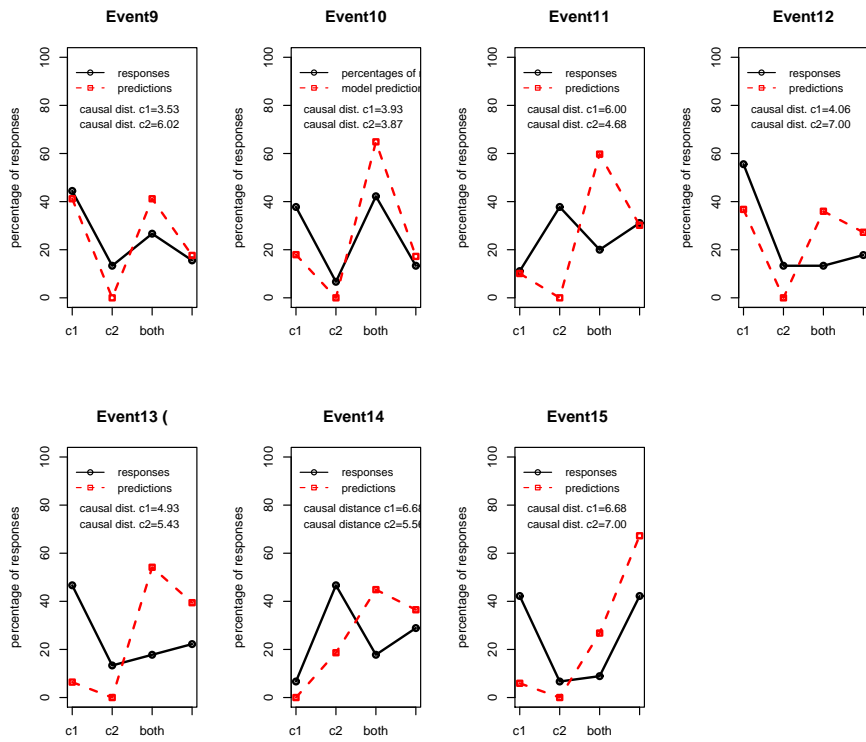


Figure D.6: Event tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

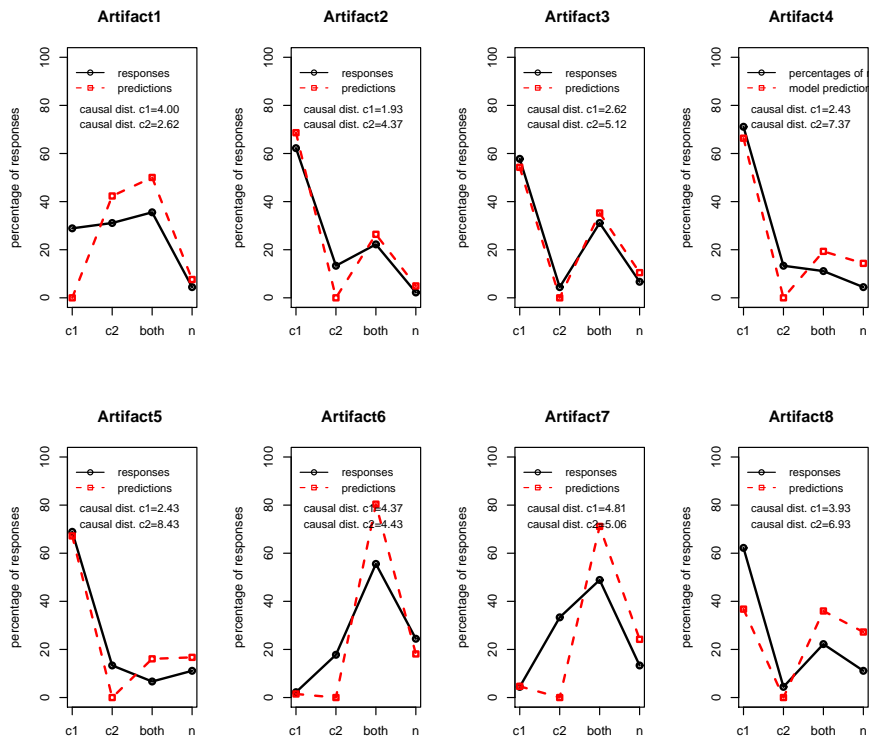


Figure D.7: Artifact tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

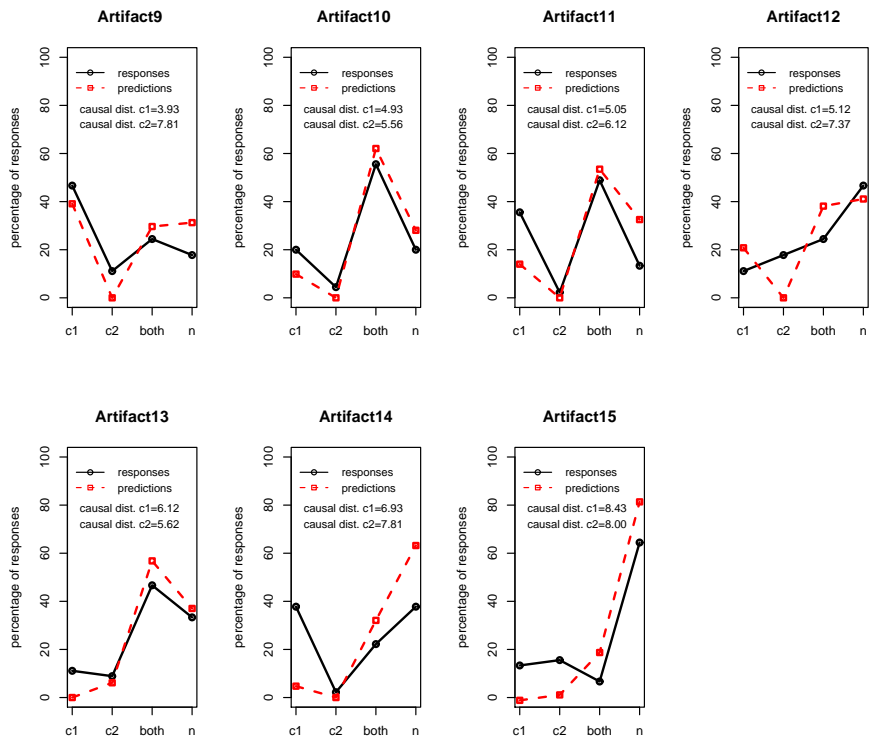


Figure D.8: Artifact tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

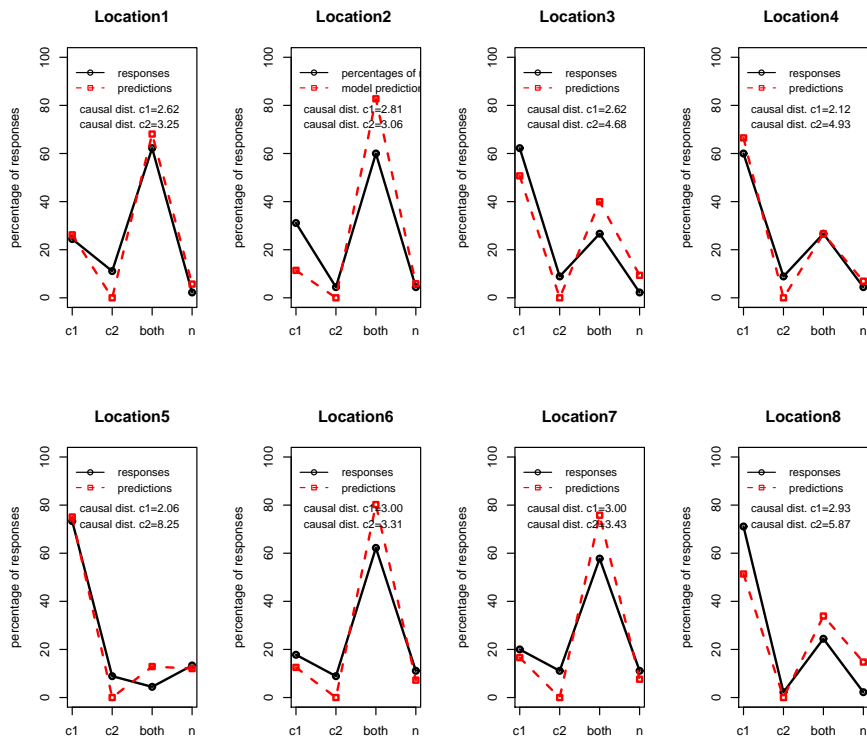


Figure D.9: Location tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

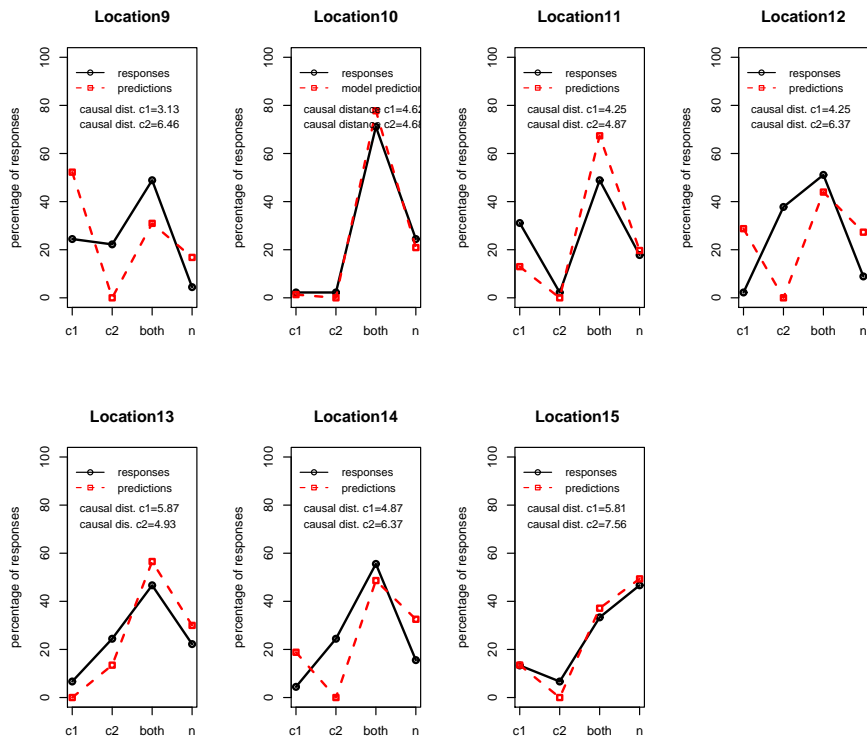


Figure D.10: Location tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

## D.4.2 Naive Causal Model Fit

Graphical representations of the Naive Causal Model fitting for the 15 trials of the five categories used in experiment 2.

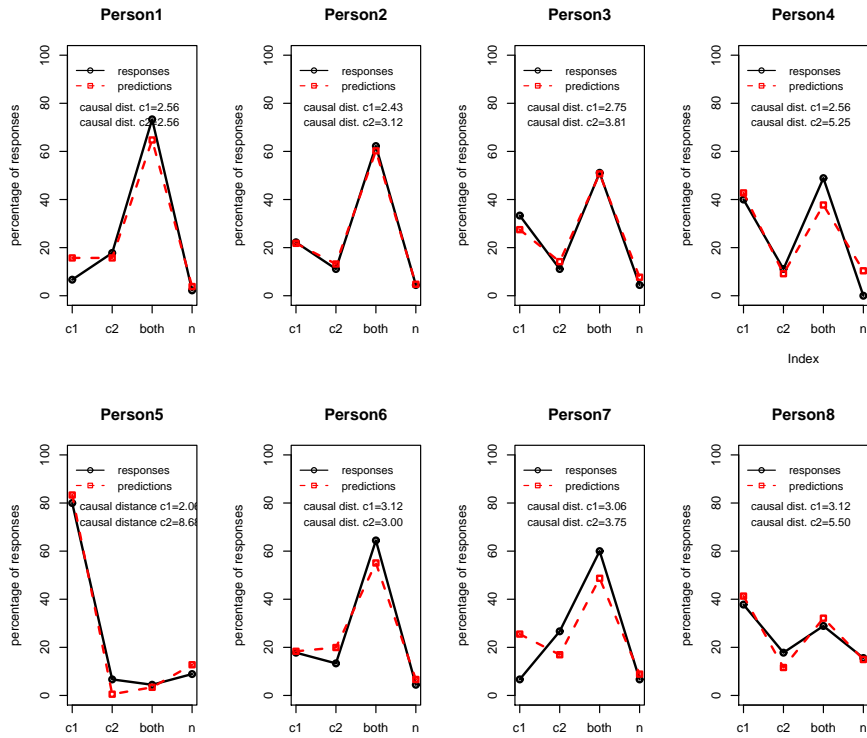


Figure D.11: Person tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

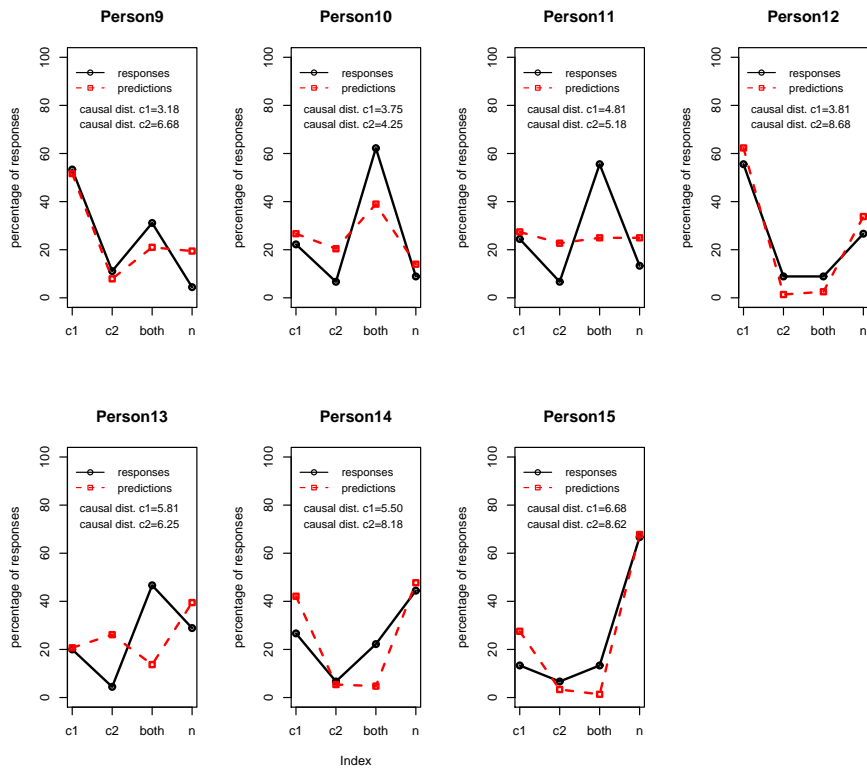


Figure D.12: Person tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.



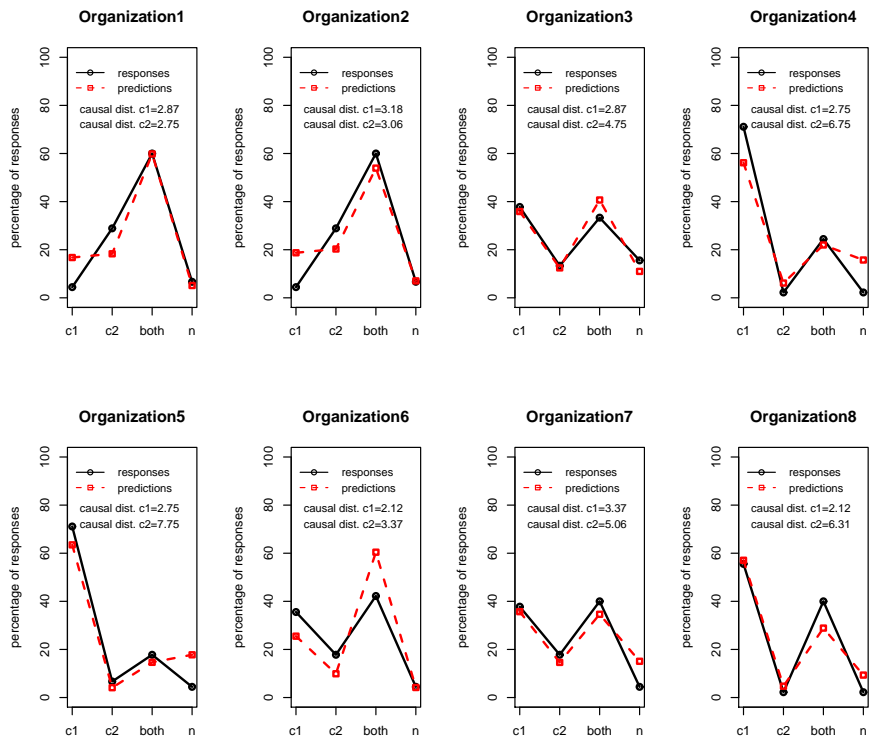


Figure D.13: Organization tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

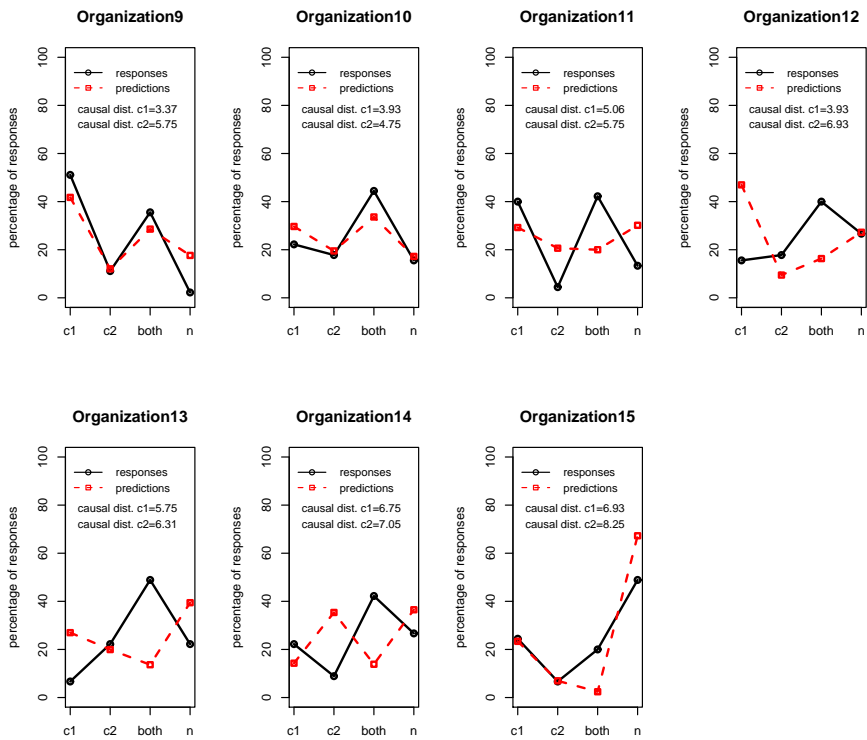


Figure D.14: Organization tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

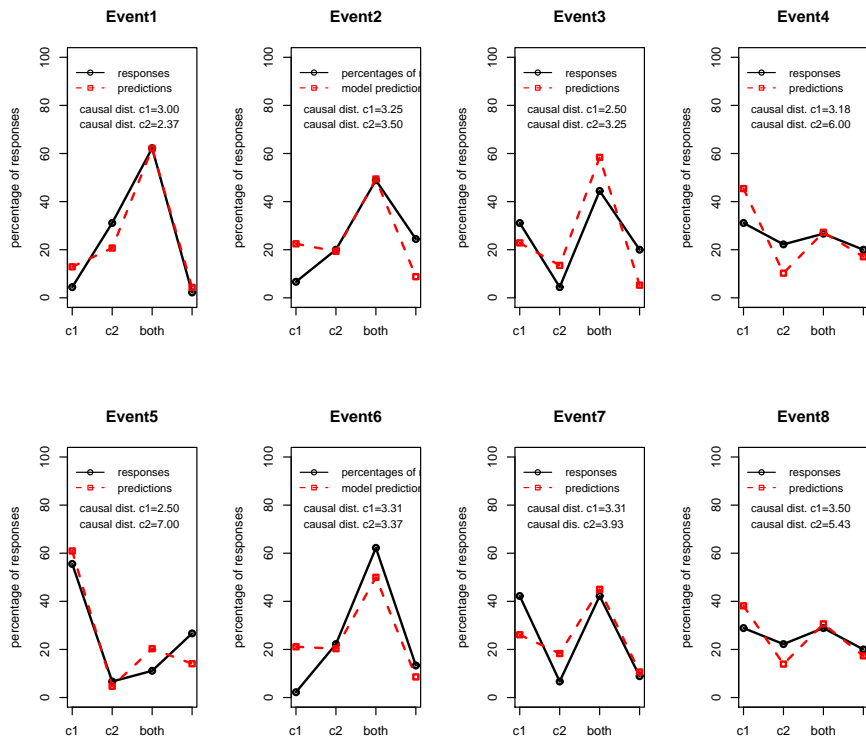


Figure D.15: Event tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

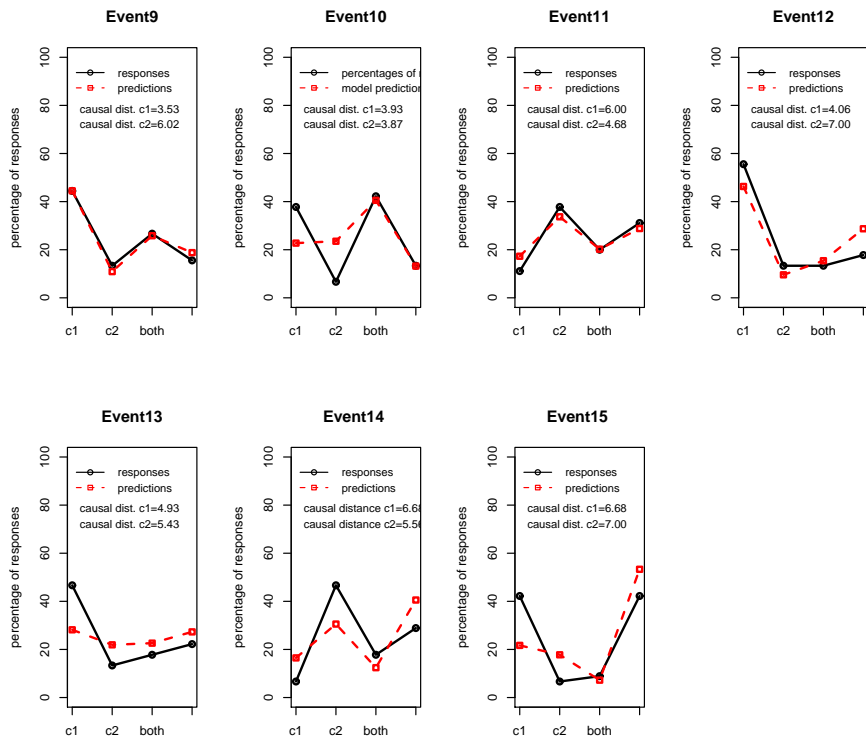


Figure D.16: Event tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

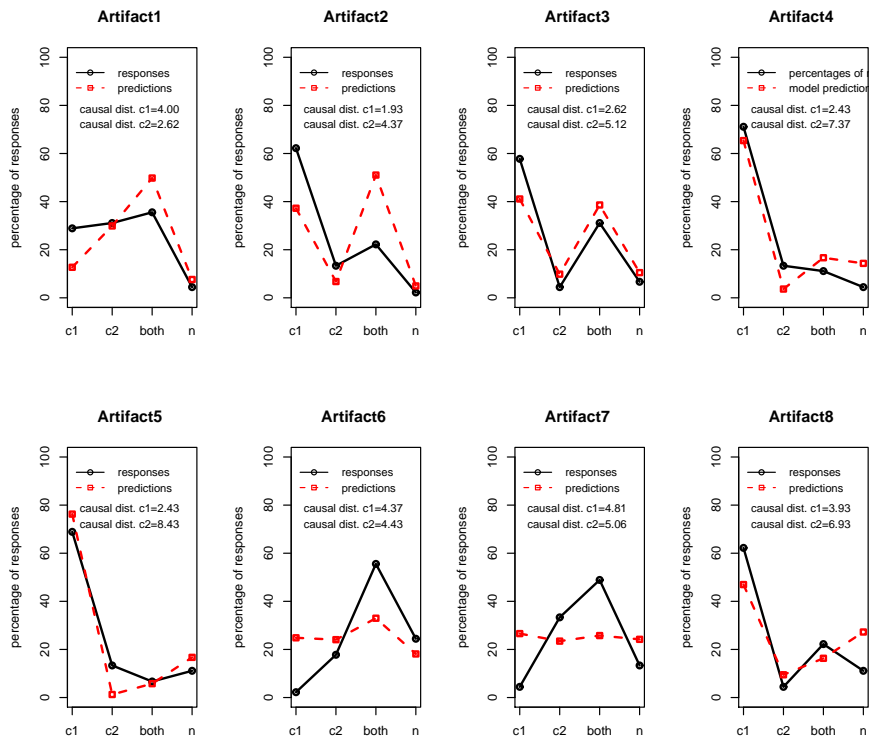


Figure D.17: Artifact tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

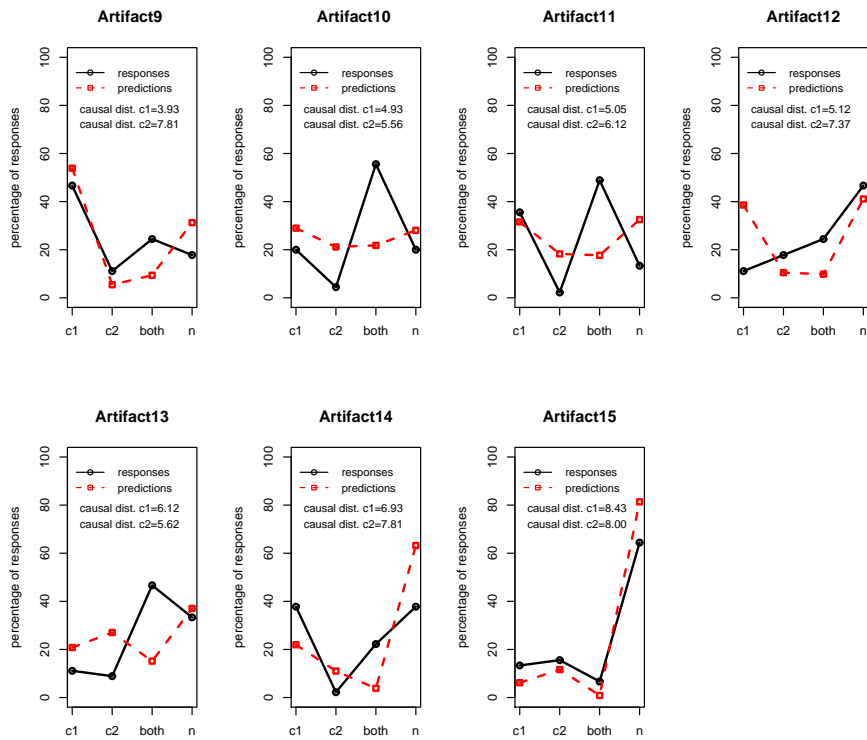


Figure D.18: Artifact tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

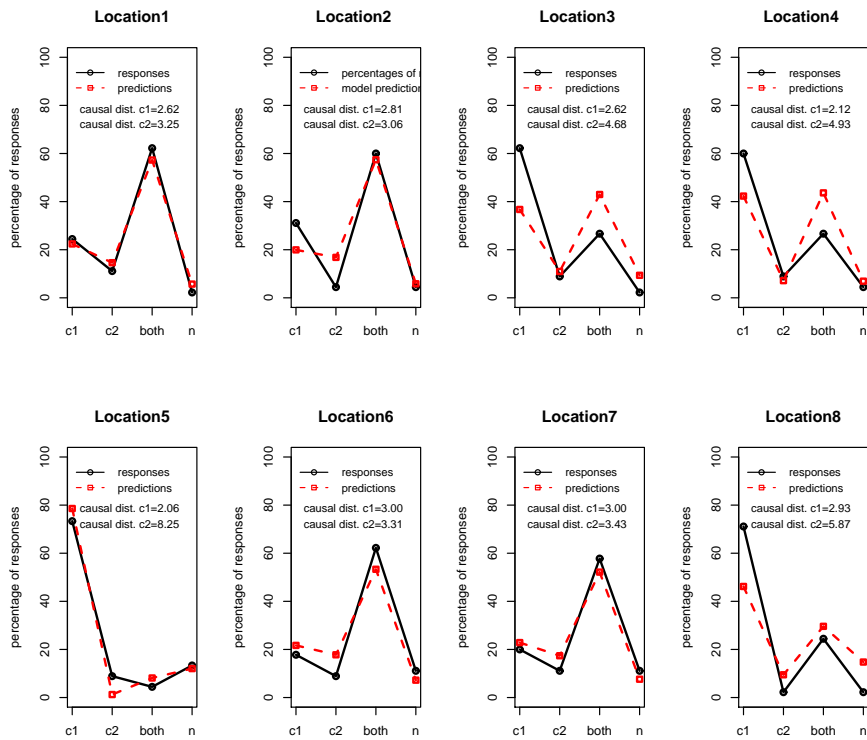


Figure D.19: Location tasks 1-8. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.

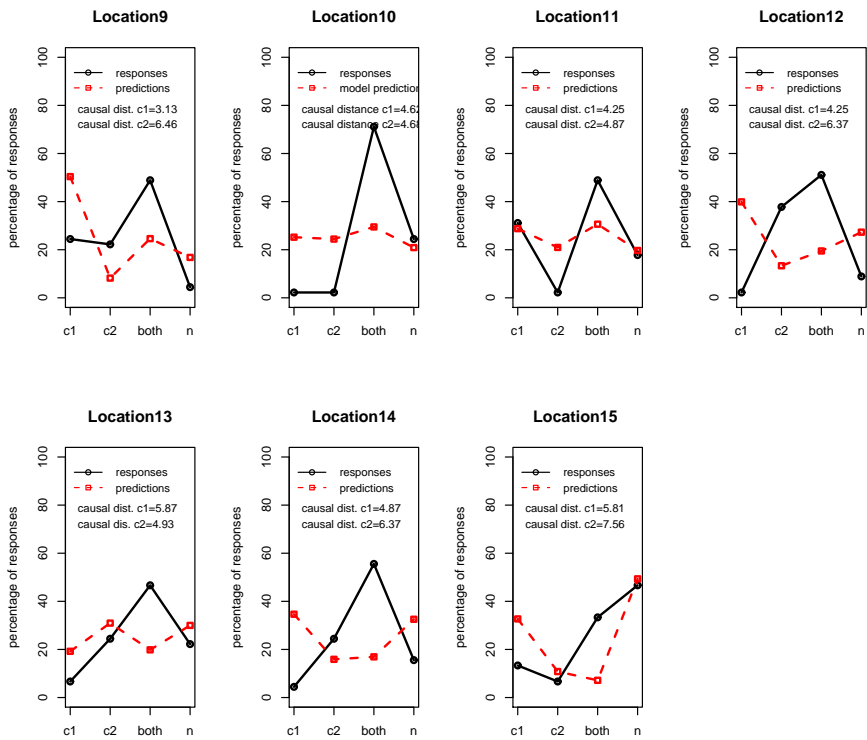


Figure D.20: Location tasks 9-15. Percentage of responses that the continuer 1, continuer 2, both continuers, or neither continuers refers to the same entity as the original description. Lines with square points are predictions from the causal continuer model.



# Appendix E

## PropLit Index

### E.1 Predicate-Entity Type Mapping used in the RDF index

In the following table we reported the mapping schema used in the advanced search module of the PropLit Index.

Predicate	Mapped in
<a href="http://www.aktors.org/ontology/portal#full-name">http://www.aktors.org/ontology/portal#full-name</a>	person
<a href="http://www.aktors.org/ontology/portal#given-name">http://www.aktors.org/ontology/portal#given-name</a>	person
<a href="http://www.aktors.org/ontology/portal#family-name">http://www.aktors.org/ontology/portal#family-name</a>	person
<a href="http://xmlns.com/foaf/0.1/name">http://xmlns.com/foaf/0.1/name</a>	person
<a href="http://xmlns.com/foaf/0.1/member_name">http://xmlns.com/foaf/0.1/member_name</a>	person
<a href="http://xmlns.com/foaf/0.1/givenname">http://xmlns.com/foaf/0.1/givenname</a>	person
<a href="http://xmlns.com/foaf/0.1/surname">http://xmlns.com/foaf/0.1/surname</a>	person
<a href="http://xmlns.com/foaf/0.1/nick">http://xmlns.com/foaf/0.1/nick</a>	person
<a href="http://xmlns.com/foaf/0.1/accountName">http://xmlns.com/foaf/0.1/accountName</a>	person
<a href="http://www.w3.org/2006/vcard/ns#given-name">http://www.w3.org/2006/vcard/ns#given-name</a>	person
<a href="http://www.w3.org/2006/vcard/ns#fn">http://www.w3.org/2006/vcard/ns#fn</a>	person
<a href="http://purl.oclc.org/NET/nknouf/ns/bibtex#hasAuthor">http://purl.oclc.org/NET/nknouf/ns/bibtex#hasAuthor</a>	person
<a href="http://rdfs.org/sioc/ns#name">http://rdfs.org/sioc/ns#name</a>	person
<a href="http://purl.org/dc/elements/1.1/creator">http://purl.org/dc/elements/1.1/creator</a>	person
<a href="http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfPerson">http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfPerson</a>	person
<a href="http://purl.uniprot.org/core/author">http://purl.uniprot.org/core/author</a>	person
<a href="http://dbpedia.org/property/artist">http://dbpedia.org/property/artist</a>	person
<a href="http://dbpedia.org/property/playername">http://dbpedia.org/property/playername</a>	person
<a href="http://www.aktors.org/ontology/portal#has-appellation">http://www.aktors.org/ontology/portal#has-appellation</a>	person
<a href="http://www.rdfabout.com/rdf/schema/ussec/officerTitle">http://www.rdfabout.com/rdf/schema/ussec/officerTitle</a>	person
<a href="http://purl.uniprot.org/core/author">http://purl.uniprot.org/core/author</a>	person
<a href="http://dbpedia.org/ontology/leaderTitle">http://dbpedia.org/ontology/leaderTitle</a>	person
<a href="http://www.geonames.org/ontology#name">http://www.geonames.org/ontology#name</a>	location
<a href="http://www.geonames.org/ontology#alternateName">http://www.geonames.org/ontology#alternateName</a>	location
<a href="http://www.geonames.org/ontology/#locationMap">http://www.geonames.org/ontology/#locationMap</a>	location
<a href="http://dbpedia.org/property/country">http://dbpedia.org/property/country</a>	location
<a href="http://dbpedia.org/property/county">http://dbpedia.org/property/county</a>	location
<a href="http://dbpedia.org/property/counties">http://dbpedia.org/property/counties</a>	location
<a href="http://dbpedia.org/property/birthPlace">http://dbpedia.org/property/birthPlace</a>	location
<a href="http://dbpedia.org/property/deathPlace">http://dbpedia.org/property/deathPlace</a>	location
<a href="http://dbpedia.org/property/region">http://dbpedia.org/property/region</a>	location

<a href="http://dbpedia.org/property/location">http://dbpedia.org/property/location</a>	location
<a href="http://dbpedia.org/property/state">http://dbpedia.org/property/state</a>	location
<a href="http://dbpedia.org/property/city">http://dbpedia.org/property/city</a>	location
<a href="http://dbpedia.org/property/place">http://dbpedia.org/property/place</a>	location
<a href="http://dbpedia.org/property/hqCity">http://dbpedia.org/property/hqCity</a>	location
<a href="http://dbpedia.org/property/province">http://dbpedia.org/property/province</a>	location
<a href="http://dbpedia.org/property/region">http://dbpedia.org/property/region</a>	location
<a href="http://dbpedia.org/property/frazioni">http://dbpedia.org/property/frazioni</a>	location
<a href="http://dbpedia.org/property/capital">http://dbpedia.org/property/capital</a>	location
<a href="http://dbpedia.org/property/citta">http://dbpedia.org/property/citta</a>	location
<a href="http://dbpedia.org/property/deathPlace">http://dbpedia.org/property/deathPlace</a>	location
<a href="http://dbpedia.org/property/adresse">http://dbpedia.org/property/adresse</a>	location
<a href="http://rdf.geospecies.org/ont/gsonontology#stateprov_name">http://rdf.geospecies.org/ont/gsonontology#stateprov_name</a>	location
<a href="http://tap.xmlns.com/data/representsPlace">http://tap.xmlns.com/data/representsPlace</a>	location
<a href="http://dbpedia.org/property/hometown">http://dbpedia.org/property/hometown</a>	location
<a href="http://www.ontoknowledge.org/oil/case-studies/Waterways">http://www.ontoknowledge.org/oil/case-studies/Waterways</a>	location
<a href="http://swat.cse.lehigh.edu/resources/onto/university.owl#city">http://swat.cse.lehigh.edu/resources/onto/university.owl#city</a>	location
<a href="http://www.w3.org/2006/vcard/ns#locality">http://www.w3.org/2006/vcard/ns#locality</a>	location
<a href="http://www.w3.org/2001/vcard-rdf/3.0#Locality">http://www.w3.org/2001/vcard-rdf/3.0#Locality</a>	location
<a href="http://xmlns.com/foaf/0.1/based_near">http://xmlns.com/foaf/0.1/based_near</a>	location
<a href="http://www.w3.org/2006/vcard/ns#country-name">http://www.w3.org/2006/vcard/ns#country-name</a>	location
<a href="http://data.linkedct.org/resource/linkedct/facility_address_country">http://data.linkedct.org/resource/linkedct/facility_address_country</a>	location
<a href="http://data.linkedct.org/resource/linkedct/facility_address_city">http://data.linkedct.org/resource/linkedct/facility_address_city</a>	location
<a href="http://www.w3.org/2002/12/cal/icaltzd#location">http://www.w3.org/2002/12/cal/icaltzd#location</a>	location
<a href="http://dbpedia.org/property/regierungsbezirk">http://dbpedia.org/property/regierungsbezirk</a>	location
<a href="http://purl.oclc.org/NET/nknouf/ns/bibtex#hasAddress">http://purl.oclc.org/NET/nknouf/ns/bibtex#hasAddress</a>	location
<a href="http://swrc.ontoware.org/ontology#address">http://swrc.ontoware.org/ontology#address</a>	location
<a href="http://www.isi.edu/webscripiter/bibtex.o.daml#address">http://www.isi.edu/webscripiter/bibtex.o.daml#address</a>	location
<a href="http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfLocation">http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfLocation</a>	location
<a href="http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfLocation">http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfLocation</a>	location
<a href="http://annotation.semanticweb.org/iswc/iswc.daml#location">http://annotation.semanticweb.org/iswc/iswc.daml#location</a>	location
<a href="http://www.isi.edu/webscripiter/bibtex.o.daml#address">http://www.isi.edu/webscripiter/bibtex.o.daml#address</a>	location
<a href="http://www.daml.org/2002/02/telephone/1/areacodes-ont#rc">http://www.daml.org/2002/02/telephone/1/areacodes-ont#rc</a>	location
<a href="http://www.radarnetworks.com/shazam#location">http://www.radarnetworks.com/shazam#location</a>	location
<a href="http://wikicompany.org/wiki/Special:URIResolver/Property-3AAddress">http://wikicompany.org/wiki/Special:URIResolver/Property-3AAddress</a>	location
<a href="http://wikicompany.org/wiki/Special:URIResolver/Property-3ARegion">http://wikicompany.org/wiki/Special:URIResolver/Property-3ARegion</a>	location
<a href="http://www.daml.ri.cmu.edu/ont/USCity.daml#name">http://www.daml.ri.cmu.edu/ont/USCity.daml#name</a>	location
<a href="http://demo.openlinksw.com/schemas/northwind#provinceName">http://demo.openlinksw.com/schemas/northwind#provinceName</a>	location
<a href="http://www.cs.cas.cz/semweb#publisher_address">http://www.cs.cas.cz/semweb#publisher_address</a>	location
<a href="http://dbpedia.org/property/state">http://dbpedia.org/property/state</a>	location
<a href="http://www.daml.org/2001/01/gedcom/gedcom#place">http://www.daml.org/2001/01/gedcom/gedcom#place</a>	location
<a href="http://demo.openlinksw.com/schemas/northwind#shipAddress">http://demo.openlinksw.com/schemas/northwind#shipAddress</a>	location
<a href="http://www.w3.org/2006/vcard/ns#street-address">http://www.w3.org/2006/vcard/ns#street-address</a>	location
<a href="http://e-tourism.deri.at/ont/e-tourism.owl#hasStreet">http://e-tourism.deri.at/ont/e-tourism.owl#hasStreet</a>	location
<a href="http://www.snee.com/ns/flights#flightFromCityName">http://www.snee.com/ns/flights#flightFromCityName</a>	location
<a href="http://www.snee.com/ns/flights#flightToCityName">http://www.snee.com/ns/flights#flightToCityName</a>	location
<a href="http://dbpedia.org/property/cityStateProperty">http://dbpedia.org/property/cityStateProperty</a>	location
<a href="http://rdf.geospecies.org/ont/gsonontology#timezone">http://rdf.geospecies.org/ont/gsonontology#timezone</a>	location
<a href="http://www.w3.org/2001/vcard-rdf/3.0#Country">http://www.w3.org/2001/vcard-rdf/3.0#Country</a>	location
<a href="http://www4.wiwiss.fu-berlin.de/factbook/ns#countryname_localshortform">http://www4.wiwiss.fu-berlin.de/factbook/ns#countryname_localshortform</a>	location
<a href="http://www4.wiwiss.fu-berlin.de/factbook/ns#countryname_locallongform">http://www4.wiwiss.fu-berlin.de/factbook/ns#countryname_locallongform</a>	location
<a href="http://www.ontoknowledge.org/oil/case-studies/Country_name">http://www.ontoknowledge.org/oil/case-studies/Country_name</a>	location
<a href="http://www.ontoknowledge.org/oil/case-studies/National_capital">http://www.ontoknowledge.org/oil/case-studies/National_capital</a>	location
<a href="http://wikicompany.org/wiki/Special:URIResolver/Property-3ARegion">http://wikicompany.org/wiki/Special:URIResolver/Property-3ARegion</a>	location
<a href="http://data.semanticweb.org/ns/swc/ontology#affiliation">http://data.semanticweb.org/ns/swc/ontology#affiliation</a>	organization
<a href="http://www.okkam.org/prefix/affiliation">http://www.okkam.org/prefix/affiliation</a>	organization
<a href="http://www.w3.org/2006/vcard/ns#organization-name">http://www.w3.org/2006/vcard/ns#organization-name</a>	organization
<a href="http://ramonantonio.net/doac/0.1/#organization">http://ramonantonio.net/doac/0.1/#organization</a>	organization
<a href="http://purl.oclc.org/NET/nknouf/ns/bibtex#hasInstitution">http://purl.oclc.org/NET/nknouf/ns/bibtex#hasInstitution</a>	organization

<a href="http://dbpedia.org/ontology/keyPersonPosition">http://dbpedia.org/ontology/keyPersonPosition</a>	organization
<a href="http://dbpedia.org/property/airline">http://dbpedia.org/property/airline</a>	organization
<a href="http://dbpedia.org/property/manufacturer">http://dbpedia.org/property/manufacturer</a>	organization
<a href="http://data.semanticweb.org/ns/swc/ontology#affiliation">http://data.semanticweb.org/ns/swc/ontology#affiliation</a>	organization
<a href="http://dbpedia.org/property/secondTeam">http://dbpedia.org/property/secondTeam</a>	organization
<a href="http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfInstitution">http://www.kisti.re.kr/isrl/ResearchRefOntology#engNameOfInstitution</a>	organization
<a href="http://dbpedia.org/property/firstTeam">http://dbpedia.org/property/firstTeam</a>	organization
<a href="http://dbpedia.org/property/fastTeam">http://dbpedia.org/property/fastTeam</a>	organization
<a href="http://dbpedia.org/property/thirdTeam">http://dbpedia.org/property/thirdTeam</a>	organization
<a href="http://dbpedia.org/property/companyName">http://dbpedia.org/property/companyName</a>	organization
<a href="http://dbpedia.org/property/clubname">http://dbpedia.org/property/clubname</a>	organization
<a href="http://dbpedia.org/property/acronyms">http://dbpedia.org/property/acronyms</a>	organization
<a href="http://www.w3.org/2006/vcard/ns#organization-unit">http://www.w3.org/2006/vcard/ns#organization-unit</a>	organization
<a href="http://www.w3.org/2006/vcard/ns#organization-name">http://www.w3.org/2006/vcard/ns#organization-name</a>	organization
<a href="http://dev.livingreviews.org/epubtk/terms#affiliation">http://dev.livingreviews.org/epubtk/terms#affiliation</a>	organization
<a href="http://ramonantonio.net/doac/0.1/#organization">http://ramonantonio.net/doac/0.1/#organization</a>	organization
<a href="http://purl.org/dc/elements/1.1/publisher">http://purl.org/dc/elements/1.1/publisher</a>	organization
<a href="http://dbpedia.org/property/employer">http://dbpedia.org/property/employer</a>	organization
<a href="http://www.aktors.org/ontology/portal#has-goals">http://www.aktors.org/ontology/portal#has-goals</a>	organization
<a href="http://www.cs.utexas.edu/users/ml/riddle#publisher">http://www.cs.utexas.edu/users/ml/riddle#publisher</a>	organization

Table E.1: Predicate-Entity Type mapping schema

## E.2 Top-50 RDF Predicates and their frequency

In the following table we reported the list of the top-50 RDF predicates and the corresponding frequencies in the original data set of RDF triples, before filtering the data set to extract predicate-literal couples. The Table gives an idea of the distribution of triples between the main ontologies today available.

<b>Predicate</b>	<b>Frequency</b>
<a href="http://dbpedia.org/property/wikilink">http://dbpedia.org/property/wikilink</a>	156,434,900
<a href="#">rdf:type</a>	143,479,200
<a href="#">rdfs:seeAlso</a>	53,852,300
<a href="#">foaf:knows</a>	35,786,400
<a href="#">foaf:nick</a>	32,979,500
<a href="#">foaf:weblog</a>	23,239,200
<a href="#">dc:title</a>	22,356,700
<a href="#">akt:has-author</a>	19,541,900
<a href="#">sioc:links_to</a>	19,228,400
<a href="#">skos:subject</a>	18,280,600
<a href="#">foaf:interest</a>	16,786,400
<a href="#">foaf:member_name</a>	14,799,800
<a href="#">rss:link</a>	14,357,800
<a href="#">foaf:holdsAccount</a>	14,038,900
<a href="#">foaf:image</a>	13,871,800
<a href="#">rss:title</a>	13,524,600
<a href="#">rdfs:label</a>	13,515,900
<a href="#">foaf:name</a>	13,179,000
<a href="#">geonames:nearbyFeatures</a>	13,128,700
<a href="#">dc:date</a>	12,519,700
<a href="#">foaf:accountName</a>	12,133,000
<a href="#">foaf:accountServiceHomepage</a>	12,068,600
<a href="#">geonames:parentFeature</a>	11,466,300
<a href="#">foaf:tagLine</a>	10,677,500
<a href="#">rss:description</a>	9,844,700

content:encoded	9,794,800
foaf:accountProfilePage	9,483,700
sioc:has_container	9,171,900
rdfs:comment	9,109,600
akt:cites-publication-reference	8,944,600
geonames:name	7,600,800
geo:lat	7,399,900
geo:long	7,341,000
http://dbpedia.org/property/wikiPageUsesTemplate	7,102,100
akt:full-name	7,100,700
dc:creator	6,987,900
geonames:featureClass	6,962,200
geonames:inCountry	6,827,100
geonames:locationMap	6,822,700
geonames:featureCode	6,822,300
owl:sameAs	6,539,300
http://dbpedia.org/property/redirect	6,451,500
foaf:homepage	6,427,100
http://dbpedia.org/property/abstract	5,750,400
foaf:img	5,562,800
http://purl.org/rss/1.0/modules/rss091#pubDate	4,814,700
foaf:page	4,653,200
dc:description	4,651,700
akt:has-title	4,310,500
akt:has-date	3,923,200

Table E.2: Top-50 RDF Predicates and their frequency in the 1 billion triple store.

# Bibliography

- [1] <http://www.l3s.de/ioannou/entityrequests.html>.
- [2] E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proc. International Conference on General WordNet*, 2002.
- [3] J. R. Anderson. Memory for information about individuals. *Memory and Cognition*, 5:430–442, 1977.
- [4] J. M. H. M. D. Anes. Roles of object-file review and type priming in visual identification within and across eye fixations. *Journal of Experimental Psychology: Human Perception and Performance*, 20:826–839, 1994.
- [5] M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Proc. Human Language Technology conference - North American chapter of the Association for Computational Linguistics*, 2003.
- [6] A. Aula and K. Nordhausen. Modeling successful performance in web searching. *Journal of the American Society for Information Science and Technology*, 57:1678 – 1693, 2006.
- [7] K. Bach. *Thought and reference*. Oxford: Clarendon Press., 1987.
- [8] R. Baillargeon and M. Graber. Where’s the rabbit?: 5.5-month-old infants’ representation of the height of a hidden object. *Cognitive Development*, 2,, 2:375–392, 1987.
- [9] C. Barry, R. A. Johnston, and L. C. Scanlan. Are faces “special” objects? Associative and semantic priming of face and object recognition and naming. *The Quarterly Journal of Experimental Psychology*, 51A:853–882, 1998.
- [10] L. W. Barsalou. Context-independent and context-dependent information in concepts. *Memory & Cognition*, 10:82–93, 1982.

- [11] L. W. Barsalou. Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. C. Collins, S. E. Gathercole, and M. A. Conway, editors, *Theories of memory*. London: Lawrence Erlbaum Associates, 1993.
- [12] L. W. Barsalou. Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Series B*, 358:1177–1187, 2003.
- [13] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 2004 ACM Conference on Research and Development in Information Retrieval (SIGIR-2004)*, Sheffield, UK, 2004.
- [14] S. M. Beitzel, E. C. Jensen, A. Chowdhury, and O. Frieder. Varying approaches to topical web query classification. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 784, 2007.
- [15] S. M. Beitzel, E. C. Jensen, O. Frieder, D. Grossman, D. D. Lewis, A. Chowdhury, and A. Kolcz. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, page 582, 2005.
- [16] T. R. Belin and D. B. Rubin. A method for calibrating false-match rates in record linkage. In *In Record Linkage - 1997: Proceedings of an International Workshop and Exposition*, 1997.
- [17] B. Belke, H. Leder, G. Harsanyi, and C. Carbon. When a picasso is a "picasso": The entry point in the identification of visual art. *Acta Psychologica*, 133:191–202, 2010.
- [18] S. Bergsma and Q. I. Wang. Learning noun phrase query segmentation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.
- [19] T. Berners-Lee. What the Semantic Web isn't but can represent. <http://www.w3.org/DesignIssues/RDFnot.html>, 1998.
- [20] T. Berners-Lee, J. A. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May, 2001.

- [21] E. Bick. A named entity recognizer for danish. In *Conference on Language Resources and Evaluation*, 2004.
- [22] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *Proc. Conference on Applied Natural Language Processing*, 1997.
- [23] S. Blok, G. Newman, and L. J. Rips. Individuals and their concepts. In W. kyoung Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff, editors, *Categorization Inside and Outside the Laboratory: Essays in Honor of Douglas L. Medin*. Washington, D.C.: American Psychological Association., 2005.
- [24] S. V. Blok, G. Newman, and L. J. Rips. Out of sorts? some remedies for theories of object concepts: A reply to rhemtulla and xu (2007). *Psychological Review*, 114:1096–1104, 2007.
- [25] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Nyu: Description of the mene named entity system as used in muc-7. In *Proc. Seventh Message Understanding Conference*, 1998.
- [26] P. Bouquet, H. Stoermer, and B. Bazzanella. An entity name system (ens) for the semantic web. In *ESWC*, pages 258–272, 2008.
- [27] P. Bouquet, H. Stoermer, and D. Giacomuzzi. OKKAM: Enabling a Web of Entities. In *i3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.*, CEUR Workshop Proceedings, ISSN 1613-0073, May 2007. online [http://CEUR-WS.org/Vol-249/submission\\_150.pdf](http://CEUR-WS.org/Vol-249/submission_150.pdf).
- [28] P. Bouquet, H. Stoermer, M. Mancioffi, and D. Giacomuzzi. Okkam: Towards a solution to the identity crisis. In *Proceedings of the Third Workshop on Semantic Web: Applications and Perspective (SWAP2006)*, Pisa, Italy, 2006.
- [29] P. Bouquet, H. Stoermer, C. Niederee, and A. Mana. Entity name system: The backbone of an open and scalable web of data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, 2008.
- [30] D. J. Brenes, D. Gayo-Avello, and K. PÃ©rez-GonzÃ¡lez. Survey and evaluation of query intent detection methods. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 1–7, Barcelona, Spain, 2009. ACM.

- [31] T. Brennen and B. Bruce. Context effects in the processing of familiar faces. *Psychological Research*, 53:296–304, 1991.
- [32] S. Brin. Extracting patterns and relations from the world wide web. In *Proc. Conference of Extending Database Technology. Workshop on the Web and Databases.*, 1998.
- [33] A. Broder. A taxonomy of web search. In *ACM Sigir Forum*, volume 36, pages 3–10, 2002.
- [34] V. Bruce. Recognizing faces. *Philosophical Transactions of the Royal Society of London*, 302:423–436, 1983.
- [35] V. Bruce and A. Young. Understanding face recognition. *British Journal of Psychology*, 77:305–327, 1986.
- [36] S. Brédart, T. Valentine, A. Calder, and L. Gassi. An interactive activation model of face naming. *Quarterly Journal Of Experimental Psychology Section A*, 48:466–86, 1995.
- [37] A. M. Burton and V. Bruce. Naming faces and naming names: Exploring an interactive activation model of person recognition. *Memory*, 1:457–480, 1993.
- [38] A. M. Burton, V. Bruce, and R. Johnston. Understanding face recognition with an interactive activation model. *British Journal of Psychology*, 81:361–380, 1990.
- [39] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, and Q. Yang. Context-aware query classification. In *SIGIR' 09, The 32nd Annual ACM SIGIR Conference*, 2009.
- [40] A. Caramazza and J. Shelton. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10:1–34, 1998.
- [41] S. Carey and F. Xu. Infants' knowledge of objects: beyond objects files and object tracking. *Cognition*, 80:179–213, 2001.
- [42] D. Carson and A. Burton. Semantic priming of person recognition: Categorical priming may be a weaker form of the associative priming effect. *Quarterly Journal Of Experimental Psychology Section A*, 54:1155–1179, 2001.



- [43] T. Cheng and K. C. Chang. Entity search engine: Towards agile best-effort information integration over the web. In *the Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 108–113, 2007.
- [44] K. G. Clark. Identity crisis. XML.com, September 11 2002. <http://www.xml.com/pub/a/2002/09/11/deviant.html>.
- [45] G. Cohen. Why is it difficult to put names to faces? *British Journal of Psychology*, 81:287–297, 1990.
- [46] G. Cohen and D. Faulkner. Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology*, 4:187–197, 1986.
- [47] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, 2003.
- [48] C. A. Colin and P. A. McMullen. Subordinate-level categorization relies on high spatial frequencies to a greater degree than basic-level categorization. *Perception & Psychophysics*, 67:354–364, 2005.
- [49] A. Collins and E. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.
- [50] A. M. Collins and M. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior*, 8:240–247, 1969.
- [51] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
- [52] A. Cucchiarelli, D. Luzi, and P. Velardi. Automatic semantic tagging of unknown proper names. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada, August 10-14 1998.
- [53] A. Damasio, H. Damasio, and G. V. Hoesen. Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology*, 32:331–341, 1982.
- [54] A. Damasio, D. Tranel, and H. Damasio. Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, 13:89–109, 1990.
- [55] M. F. Damian and R. A. Rahman. Semantic priming in the name retrieval of objects and famous faces. *British Journal of Psychology*, 94:517–527, 2003.

- [56] E. H. de Haan, A. Young, and F. Newcombe. Face recognition without awareness. *Cognitive Neuropsychology*, 4:385–415, 1987.
- [57] R. Delbru. Methodology for Searching Entities on the Web. In *ESWC 2008 Ph. D. Symposium*, page 26. Citeseer, 2008.
- [58] J. Driver and G. C. Baylis. Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognitive Psychology*, 31:248–306, 1996.
- [59] J. Du, Z. Zhang, J. Yan, Y. Cui, and Z. Chen. Using search session context for named entity recognition in query. In *SIGIR'10*, Geneva, Switzerland, July 19-23 2010.
- [60] S. Dubois, B. Rossion, C. Schiltz, J. M. Bodart, C. Michel, R. Bruyer, and M. Crommelinck. Effect of familiarity on the processing of human faces. *NeuroImage*, 9:278–289, 1999.
- [61] M. Dummett. *Frege: Philosophy of language*. Harvard University Press, 1973.
- [62] A. W. Ellis, A. Young, and E. Critchley. Loss of memory for people following temporal lobe damage: A case study. *Brain*, 112:1469–1483, 1989.
- [63] F. M. Engst, M. Martín-Loeches, and W. Sommer. Memory systems for structural and semantic knowledge of faces and buildingg. *Brain Research*, 1124:70–80, 2006.
- [64] M. J. Farah, K. L. Levinson, and K. L. Klein. Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33:661–674, 1995.
- [65] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of American Statistical Association*, 1969:1183–1210, 64.
- [66] L. Ferrand and B. New. Semantic and associative priming in the mental lexicon. In P. Bonin, editor, *Mental lexicon: some words to talk about words*. Hauppauge, NY: Nova Science, 2003.
- [67] M. Fleischman. Automated subcategorization of named entities. In *Conference of the European Chapter of Association for Computational Linguistic.*, 2001.
- [68] M. Fleischman and E. Hovy. Fine grained classification of named entities. In *Conference on Computational Linguistics*, 2002.

- [69] B. M. Flude, A. W. Ellis, and J. Kay. Face processing and name retrieval in an anomic aphasic: Names are stored separately from semantic information about familiar people. *Brain and Cognition*, 11:60–72, 1989.
- [70] J. Fodor. *Modularity of Mind*. Cambridge. Cambridge: MIT Press, 1983.
- [71] G. Frege. *The Foundations of Arithmetic*. Oxford: Blackwell, 1884/1959.
- [72] R. Fukatsu, T. Fujii, T. Tsukiura, A. Yamadori, and T. Otsuki. Proper name anomia after left temporal lobectomy: a patient study. *Neurology*, 23:1096–1099, 1999.
- [73] R. J. B. J. G. and Schmolze. An overview of the kl-one knowledge representation system. *Cognitive Science*, 9:171–216, 1985.
- [74] G. Gainotti. Different patterns of famous people recognition disorders in patients with right and left anterior temporal lesions: A systematic review. *Neuropsychologia*, 45:1591–1607, 2007.
- [75] G. Gainotti, F. Ciaraffa, M. C. Silveri, and C. Marra. Mental representation of normal subjects about the sources of knowledge in different semantic categories and unique entities. *Neuropsychology*, 23:803–812, 2009.
- [76] R. Gaizauskas, K. Humphreys, H. Cunningham, and Y. Wilks. University of sheffield: description of the lasie system as used for muc-6. In *MUC6*, pages 207–220, 1995.
- [77] M. Ganesh, J. Srivastava, and T. Richardson. Mining entityidentification rules for database integration. In *Proceedings of the KDD*, 1996.
- [78] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Lecture Notes In Computer Science, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, volume 2473, 2002.
- [79] I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3:191–197, 2000.
- [80] I. Gauthier and M. Tarr. Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37:1673–1682, 1997.
- [81] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore. Activation of the middle fusiform “face area” increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2:568–573, 1999.

- [82] I. Gauthier, M. J. Tarr, J. Moylan, P. Skudlarski, J. C. Gore, and A. W. Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, 12:495–504, 2000.
- [83] S. Gelman and M. Taylor. How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, 55:1535–1540, 1984.
- [84] V. Gentileschi, S. Sperber, and H. Spinnler. Crossmodal agnosia for familiar people as a consequence of right infero-polar temporal atrophy. *Cognitive Neuropsychology*, 18:439–463, 2001.
- [85] H. Glaser, A. Jaffri, and I. C. Millard. Managing co-reference on the semantic web. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.*, 2009.
- [86] P. F. Gontijo, J. Rayman, S. Zhang, and E. Zaide. How brand names are special: brands, words, and hemispheres. *Brain and Language*, 82:327–343, 2002.
- [87] I. Good. *Probability and the Weighing of Evidence*. London., Charles Griffin & Co., 1950.
- [88] R. Gordon and D. Irwin. The role of physical and conceptual properties in preserving object continuity. *Learning, Memory, and Cognition*, 26:136–150, 2000.
- [89] R. D. Gordon and D. E. Irwin. What's in an object file? evidence from priming studies. *Perception & Psychophysics*, 58:1260–1277, 1996.
- [90] M. Gorno-Tempini and C. Price. Identification of famous faces and buildings: a functional neuroimaging study of semantically unique items. *Brain*, 24:2087–2097, 2001.
- [91] M. Gorno-Tempini, C. Price, O. Josephs, R. Vandenberghe, S. Cappa, and N. K. et al. The neural systems sustaining face and proper-name processing. *Brain*, 121:2103–2118, 1998.
- [92] T. J. Grabowski, H. Damasio, D. Tranel, L. L. B. Ponto, R. D. Hichwa, and A. R. Damasio. A role for left temporal pole in the retrieval of words for unique entities. *Human Brain Mapping*, 13:199–212, 2001.
- [93] K. Grill-Spector and N. Kanwisher. As soon as you know it is there, you know what it is. *Psychological Science*, 16:152–160, 2005.

- [94] R. Grishman and B. Sundheim. Message understanding conference - 6: A brief history. In *International Conference on Computational Linguistics.*, 1996.
- [95] R. Guha and A. Garg. *Disambiguating People in Search*. Stanford University, 2004.
- [96] R. Guillén. Geoparsing web queries. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum*, 2008.
- [97] J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval SIGIR 09 (2009)*, 2009.
- [98] D. G. Hall, B. C. Veltkamp, and W. J. Turkel. Children’s and adults’ understanding of proper namable things. *First Language*, 24:5–32, 2004.
- [99] G. Hall. How children learn common nouns and proper names. In J. Macnamara and G. Reyes, editors, *The Logical Foundations of Cognition*. Oxford:Oxford University Press., 1994.
- [100] H. Halpin. Identity, reference and meaning on the web. In *Proceedings of the Workshop on Identity, Meaning and the Web (IMW06) at WWW2006, Edinburgh, Scotland*, 2006.
- [101] H. Halpin and V. Presutti. An ontology of resources: Solving the identity crisis. In *Proceedings of 6th Annual European Semantic Web Conference ESWC2009*, pages 121–140. Research Studies Press/Wiley, June 2009.
- [102] J. Hanley and E. Cowell. The effects of different types of retrieval cues on the recall of names of famous faces. *Memory and Cognition*, 16:545–555, 1988.
- [103] R. Hanley, A. W. Young, and N. A. Pearson. Defective recognition of familiar people. *Cognitive Neuropsychology*, 6:179–210, 1989.
- [104] J. V. Haxby, B. Horwitz, L. G. Ungerleider, J. M. Maisog, P. Pietrini, and C. L. Grady. The functional organization of human extrastriate cortex: A pet-rcbf study of selective attention to faces and locations. *The Journal of Neuroscience*, 74:6336–6353, 1994.
- [105] X. He, J. Yan, J. Ma, N. Liu, and Z. Chen. Query topic detection for reformulation. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

- [106] A. Hillis and A. Caramazza. Category specific naming and comprehension impairment: A double dissociation. *Brain*, 114:2081–2094, 1991.
- [107] E. Hirsch. *The Concept of Identity*. Oxford University Press, 1982.
- [108] J. Hodgson. Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, 6:169–205, 1991.
- [109] C. Hoelscher. How internet experts search for information on the web. Paper presented at the World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL., 1998.
- [110] G. Humphreys, M. Riddoch, and P. Quinlan. Cascade processes in picture identification. *Cognitive Neuropsychology*, 5:67–103, 1988.
- [111] K. Hutchison. Is semantic priming due to association strength or featural overlap? a micro-analytic review. *Psychonomic Bulletin and Review*, 10:785–813, 2003.
- [112] R. Jackendoff. *Semantics and Cognition*. Cambridge, Mass.: MIT Press, 1983.
- [113] B. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52:235–246, 2001.
- [114] B. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227, 2000.
- [115] R. Jeshion. The significance of names. *Mind and Language*, 24:372–405, 2009.
- [116] R. Jeshion. *New Essays on Singular Thought*. Oxford, Oxford University Press., 2010.
- [117] G. S. R. Job, M. Miozzo, S. Zago, and G. Marchiori. Category specific form-knowledge deficits in a patient with herpes simplex virus encephalitis. *Journal of Clinical and Experimental Neuropsychology*, 15:280–299, 1993.
- [118] K. E. Johnson and C. B. Mervis. Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, 126:248–277, 1997.
- [119] P. Jolicoeur, M. Gluck, and S. Kosslyn. Pictures and names: Making the connection. *Cognitive Psychology*, 16:243–275, 1984.

- [120] D. Kahneman and A. Treisman. Changing views of attention and automaticity. In R. Parasuraman and D. R. Davies, editors, *Varieties of attention*, pages 29–62. Orlando: Academic Press., 1984.
- [121] D. Kahneman, A. Treisman, and B. J. Gibbs. The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, 24:175–219, 1992.
- [122] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.
- [123] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17:4302–4311, 1997.
- [124] L. Kartsounis and T. Shallice. Modality specific semantic knowledge loss for unique items. *Cortex*, 32:109–119, 1996.
- [125] N. Katz, E. Baker, and J. Macnamara. What’s in a name? a study of how children learn common and proper names. *Child Development*, 45:469–473, 1974.
- [126] M. Kiefer. Repetition-priming modulates category-related effects on event-related potentials: further evidence for multiple cortical semantic systems. *Journal of Cognitive Neuroscience*, 17:199–211, 2005.
- [127] S. Kripke. *Naming and Necessity*. Oxford, Basil Blackwell, 1980.
- [128] S. Kripke. *Naming and Necessity*. Basil Blackwell, Boston, 1980.
- [129] J. K. Kruschke. Alcové: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44, 1992.
- [130] B. Laeng, A. Zarrinpar, and S. M. Kosslyn. Do separate processes identify objects as exemplars versus members of basic-level categories? evidence from hemispheric specialization. *Brain and Cognition*, 53:15–27, 2003.
- [131] K. Lawlor. *New Thoughts about Old Things: Cognitive Policies as the Ground of Singular Concepts*. Garland Publishing, 2001.
- [132] S. Lawrence, C. L. Giles, and K. D. Bollacker. Autonomous citation matching. In *Proceedings of the Third International Conference on Autonomous Agents*. ACM Press, 1999.

- [133] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, page 400. ACM, 2005.
- [134] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *Proceedings of the 12th international conference on World Wide Web*, 2003.
- [135] A. M. Leslie, F. Xu, P. D. Tremoulet, and B. J. Scholl. Indexing and the object concept: developing "what" and "where" systems. *Trends in Cognitive Science*, 2:10–18, 1998.
- [136] C. L. Leveroni, M. Seidenberg, A. R. Mayer, L. A. Mead, J. R. Binder, and S. M. Rao. Neural systems underlying the recognition of familiar and newly learned faces. *The Journal of Neuroscience*, 20:878–886, 2000.
- [137] J. C. Liittschwager. Children's reasoning about identity across transformations. *Dissertation Abstracts International*, 55 (10), 4623B. (UMI No. 9508399), 1995.
- [138] E. P. Lim, J. Srivastava, S. Prabhakar, and J. Richardson. Entity identification in database integration. In *Proceedings of the Ninth International Conference on Data Engineering, Los Alamitos, Ca., USA*, pages 294 – 301. IEEE Computer Society Press, 1993.
- [139] J. Locke. *An Essay Concerning Human Understanding*. New York: Prometheus Books, 1995.
- [140] L. Lombardi and G. Sartori. Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, 57:101–125, 2007.
- [141] B. Love and S. Sloman. Mutability and theory determinants of conceptual transformability. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, 654-659. Pittsburgh, PA*, 1995.
- [142] E. J. Lowe. *Kinds of Being: A Study of Individuation, Identity, and the Logic of Sortal Terms*. Oxford: Blackwell., 1989.
- [143] F. Lucchelli and E. de Renzi. Proper name anomia. *Cortex*, 28:221–230, 1992.
- [144] S. Lupker. Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, 23:709–733, 1984.
- [145] J. Macnamara. *Names for things*. Cambridge, MA: MIT Press, 1982.



- [146] J. Macnamara. *A border dispute: The place of logic in psychology*. Cambridge, MA: MIT Press, 1986.
- [147] E. Maguire, C. Frith, and L. Cipolotti. Distinct neural systems for the encoding and recognition of topography and faces. *NeuroImage*, 13:743–750, 2001.
- [148] C. J. Marsolek. Abstract visual-form representations in the left cerebral hemisphere. *Journal of Experimental Psychology: Human Perception and Performance*, 21:375–386, 1995.
- [149] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. Wonderweb deliverable d18 ontology library (final). Technical report, Laboratory For Applied Ontology - ISTC-CNR, 2003.
- [150] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira. An introduction to the syntax and content of cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering.*, Stanford, March 2006.
- [151] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of IJCAI Workshop on Information Integration on the Web*, pages 79–86, 2003.
- [152] L. McClelland and D. E. Rumelhart. An interactive activation model of the effect of context in perception. part 1. an account of basic findings. *Psychological Review*, 88:375–406, 1981.
- [153] E. McKone, N. Kanwisher, and B. Duchaine. Can generic expertise explain special processing for faces? *Trends in Cognitive Science*, 11:8–15, 2006.
- [154] J. E. McNeil and E. K. Warrington. Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology Section A*, 46:1–10, 1993.
- [155] A. McNeill and A. Burton. The locus of semantic priming effects in person recognition. *The Quarterly Journal of Experimental Psychology*, 55A:1141–1156, 2002.
- [156] K. McRae and S. Boisvert. Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:558–572, 1998.

- [157] K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559, 2005.
- [158] K. McWeeny, A. Young, D. Hay, and A. Ellis. Putting names to faces. *British Journal of Psychology*, 78:143–149, 1987.
- [159] D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological Review*, 85, 207-238., 85:207–238, 1978.
- [160] C. Mervis and E. Rosch. Categorization of natural objects. *Annual Review of Psychology*, 32:89–115, 1981.
- [161] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.
- [162] L. Miller, D. Caine, and J. Watson. A role for the thalamus in memory for unique entities. *Neurocase*, 9:504–514, 2003.
- [163] S. R. Mitroff, B. J. Scholl, and N. S. Noles. Object files can be purely episodic. *Perception*, 36:1730–1735, 2007.
- [164] A. Monge and C. Elkan. The field-matching problem: algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.
- [165] H. Moss, R. Ostrin, L. Tyler, and W. Marslen-Wilson. Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21:863–883, 1995.
- [166] B. B. Murdock. A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89:609–626, 1982.
- [167] K. Nakamura, R. Kawashima, N. Sato, A. Nakamura, M. Sugiura, and T. K. et al. Functional delineation of the human occipito-temporal areas related to face and scene processing: A pet study. *Brain*, 123:1903–1912, 2000.
- [168] K. Nakayama, Z. J. He, and S. Shimojo. Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. K. . D. N. Osherson, editor, *An invitation to cognitive science: Visual cognition*. Cambridge, MA: MIT Press, 1995.

- [169] H. Newcombe, J. Kennedy, S. Axford, and A. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.
- [170] I. Niles and A. Pease. Towards a standard upper ontology. In C. Welty and B. Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, October 17-19, 2001.
- [171] H. Noonan. The closest continuer theory of identity. *Inquiry*, 28:195–229, 1985.
- [172] R. M. Nosofsky and M. K. Johansen. Exemplar-based accounts of multiple-system phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7:375–402, 2000.
- [173] N. Noy and C. Hafner. The state of the art in ontology design. *AI Magazine*, 18:53–74, 1997.
- [174] R. Nozick. *Philosophical explanations*. Cambridge, MA: MIT Press, 1981.
- [175] A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo. Restructuring wordnet’s top-level: The ontoclean approach. In *n K. Simov (ed.) Workshop Proceedings of OntoLex’2, Ontologies and Lexical Knowledge Bases*, Las Palmas, Spain, May 27, 2002.
- [176] Y. Otsuka, K. Suzuki, T. Fujii, R. Miura, K. Endo, H. Kondo, and A. Yamadori. Proper name anomia after left temporal subcortical hemorrhage. *Cortex*, 41:39–47, 2005.
- [177] M. Paşca. Weakly-supervised discovery of named entities using web search queries. In *CIKM ’07*, 2007.
- [178] M. Paşca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching the world wide web of facts-step one: The one-million fact extraction challenge. In *Proc. National Conference on Artificial Intelligence*, 2006.
- [179] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *In Advances in Neural Information Processing Systems 15*. MIT Press, 2003.
- [180] M. Perea and E. Rosa. The effects of associative and semantic priming in the lexical decision task. *Psychological Research*, 66:180–194, 2002.
- [181] J. Perry. *Knowledge, possibility, and consciousness*. Cambridge, MA: MIT Press., 2001.

- [182] M. A. Peterson and J. H. Kim. On what is bound in figures and grounds. *Visual Cognition*, 8:329–348, 2001.
- [183] D. C. Plaut. Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19:603–639, 2002.
- [184] M. Posner and R. Mitchell. Chronometric analysis of classification. *Psychological Review*, 74:392–409, 1967.
- [185] H. T. Pu, S. L. Chuang, and Y. Yang. Subject categorization of query terms for exploring web users’ search interests. *Journal of the American Society for Information Science and Technology*, 53:617 – 630, 2002.
- [186] A. Puce, T. Allison, M. Asgari, J. C. Gore, and G. McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, 16:5205–5215, 1996.
- [187] Z. W. Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80:127–158, 2001.
- [188] E. D. Renzi. Current issues in prosopagnosia. In H. D. Ellis, M. A. Jeevesand, and F. Newcombe, editors, *Aspects of Face Processing*. Dordrecht: Martinus Nijhoff, 1986.
- [189] E. D. Renzi, P. Faglioni, D. Grossi, and P. Nichelli. Apperceptive and associative forms of prosopagnosia. *Cortex*, 27:213–221, 1991.
- [190] M. Rhemtulla and F. Xu. Sortal concepts and causal continuity: Comment on rips, blok, and newman. *Psychological Review*, 114:1087–1095, 2007.
- [191] S. Rieh and H. Xu. Patterns and sequences of multiple query reformulations in web searching: A preliminary study. In *Proceedings of the 64th ASIST Annual Meeting*, 2001.
- [192] L. J. Rips, S. Blok, and G. Newman. Tracing the identity of objects. *Psychological Review*, 113:1–30, 2006.
- [193] L. J. Rips, E. J. Shoben, and E. E. Smith. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12:1–20, 1973.
- [194] K. M. Risvik, T. Mikolajewski, and P. Boros. Query segmentation for web search. In *The Twelfth International World Wide Web Conference (WWW)*, 2003.

- [195] T. T. Rogers and D. C. Plaut. Connectionist perspectives on category-specific deficits. In E. Forde and G. Humphreys, editors, *Category specificity in brain and mind*. Hove, UK: Psychology Press, 2002.
- [196] E. Rosch. Cognition and categorization. In H. Ellis, M. Jeeves, F. Newcombe, and A. Young, editors, *Principle of categorization*, pages 27–48. Hillsdale, NJ: Erlbaum, 1975.
- [197] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Hillsdale NJ: Erlbaum, 1978.
- [198] E. Rosch, C. Mervis, W. Gray, D. Johnson, and B.-P. Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.
- [199] E. Rosch and C. B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.
- [200] E. H. Rosch. On the internal structure of perceptual and semantic categories. In T. E. Moore, editor, *Cognitive development and the acquisition*. New York: Academic Press, 1973.
- [201] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, 2004.
- [202] N. C. M. Ross and D. Wolfram. End user searching on the internet: An analysis of term pair topics submitted to the excite search engine. *Journal of the American Society for Information Science*, 51:949–958, 2000.
- [203] P. Rotshtein, R. N. A. Henson, A. Treves, J. Driver, and R. J. Dolan. Morphing marilyn into maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, 8:107–113, 2005.
- [204] R. Rumiati, G. Humphreys, M. Riddoch, and A. Bateman. Visual object agnosia without prosopagnosia or alexia: Evidence for hierarchical theories of visual recognition. *Visual Cognition*, 1:181–225, 1994.
- [205] O. Sacks. *The Man Who Mistook His Wife for a Hat*. London: Duckworth; New York: Summit Books, 1985.
- [206] G. Salton and B. C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–123, 1988.
- [207] G. Sartori and L. Lombardi. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16:439–452, 2004.

- [208] G. Sartori, L. Lombardi, and L. Mantiuzzi. Semantic relevance best predicts normal and abnormal name retrieval. *Neuropsychologia*, 43:754–770, 2005.
- [209] G. Sartori, G. Negri, I. Mariani, and S. Prioni. Relevance of semantic features and category specificity. *Cortex*, 40:191–193, 2004.
- [210] C. Schiltz and B. Rossion. Faces are represented holistically in the human occipito-temporal cortex. *Neuroimage*, 32:1385–1394, 2006.
- [211] B. Scholl and Z. Pylyshyn. Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, 38:259–290, 1999.
- [212] B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80:1–46, 2001.
- [213] M. Schweich, M. van der Linden, S. Bredart, R. Bruyer, B. Nelles, and J.-P. Schils. Daily-life difficulties in person recognition reported by young and elderly subjects. *Applied Cognitive Psychology*, 6:161–172, 1992.
- [214] S. R. Schweinberger, A. M. Burton, and S. W. Kelly. Priming the access to names of famous faces. *British Journal of Psychology*, 92:303–307, 2001.
- [215] E. Sciore, M. Siegel, and A. Rosenthal. Using semantic values to facilitate interoperability among heterogeneous information systems. *ACM Transactions on Database Systems*, 19:254 – 290, 1994.
- [216] C. Semenza. The neuropsychology of proper names. *Mind & Language*, 24:347–369, 2009.
- [217] C. Semenza, M. Zettin, and F. Borgo. Names and identification: An access problem. *Neurocase*, 4:45–53, 1998.
- [218] J. Sergent, S. Ohta, and B. MacDonald. Functional neuroanatomy of face and object processing. a positron emission tomography study. *Brain*, 115:15–36, 1992.
- [219] J. Shelton and R. Martin. How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18:1191–1210, 1992.
- [220] D. Shen, R. Pan, J. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q<sub>2</sub>/C@UST: our winning solution to query classification in KDD-CUP 2005. *SIGKDD Explor. Newsl.*, 7(2):100–110, 2005.
- [221] D. Shen, J.-T. Sun, and a. Z. C. Q. Yang. Building bridges for web query classification. In *SIGIR '06*, 131-138.

- [222] D. Shen, T. Walker, Z. Zheng, Q. Yang, and Y. Li. Personal name classification in web queries. In *Proceedings of the international conference on Web search and web data mining*, 2008.
- [223] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. SIGIR Forum, 1999.
- [224] P. Singla and P. Domingos. Entity resolution with markov logic. In *Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA*, 2006.
- [225] S. A. Sloman, B. C. Love, and W. kyoung Ahn. Feature centrality and conceptual coherence. *Cognitive Science*, 22:189–228, 1998.
- [226] E. Smith and D. L. D.L. Medin. *Categories and concepts*. Cambridge, MA: Harvard University Press., 1981.
- [227] W. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27:521–544, 2001.
- [228] C. M. Sorrentino. Children and adults represent proper names as referring to unique individuals. *Developmental Science*, 4:399–407, 2001.
- [229] E. S. Spelke. Preferential looking methods as tools for the study of cognition in infancy. In Gottlieb and Krasnegor, editors, *Measurement of Audition and Vision in the First Year of Postnatal Life*. Nonwood, NJ: Ablex, 1985.
- [230] E. S. Spelke, R. Kestenbaum, D. Simons, and D. Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13:113–142, 1995.
- [231] A. Spink and B. J. Jansen. *Web search: Public searching of the Web*. Dordrecht: Kluwer, 2004.
- [232] A. Spink, S. Ozmutlu, H. Ozmutlu, and B. J. J. B. J. U.s. versus european web searching trends. ACM SIGIR Forum, 2002.
- [233] R. S.Taylor. Process of asking questions. *American Documentation*, 13:391–396, 1962.
- [234] S. Stevenage and H. Lewis. Understanding person acquisition using an interactive activation and competition network. *Visual Cognition*, 9:839–867, 2002.

- [235] H. Stoermer. *Okkam Enabling Entity-centric Information Integration in the Semantic Web*. PhD thesis, University of Trento, 2008.
- [236] A. Stone and T. Valentine. The categorical structure of knowledge for famous people (and a novel application of centre-surround theory). *Cognition*, 104:535–564, 2007.
- [237] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceedings of the World Wide Conference*, 2008.
- [238] J. Tanaka and M. Farah. Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology A.*, 46:225–245, 1993.
- [239] J. Tanaka and I. Gauthier. Expertise in object and face recognition. In R. Goldstone, P. Schyns, and D. Medin, editors, *Psychology of Learning and Motivation, Mechanisms of Perceptual Learning*, pages 83–125. San Diego: Academic Press, 1997.
- [240] J. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23:457–482, 1991.
- [241] J. W. Tanaka. The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology: General*, 130:534–543, 2001.
- [242] M. J. Tarr and Y. D. Cheng. Learning to see faces and objects. *Trends in Cognitive Science*, 7:23–30, 2003.
- [243] R. S. Taylor. Question-negotiation and information-seeking in libraries. *College and Research Libraries*, 29:178–194, 1968.
- [244] S. Tejada, C. A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26:607–633, 2001.
- [245] S. A. Thompson, K. S. Grahamb, G. Williams, K. Patterson, N. Kapur, and J. R. Hodges. Dissociating person-specific from general semantic knowledge: roles of the left and right temporal lobes. *Neuropsychologia*, 42:359–370, 2004.
- [246] D. Tranel. Impaired naming of unique landmarks is associated with left temporal polar damage. *Neuropsychology*, 20:1–10, 2006.
- [247] D. Tranel, H. Damasio, and A. Damasio. A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, 35:1319–1327, 1997.



- [248] A. Treisman. Perceiving and re-perceiving objects. *American Psychologist*, 47:862–875, 1992.
- [249] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.
- [250] T. Valentine, S. Bredart, R. Lawson, and G. Ward. What’s in a name? access to information from people’s names. *European Journal of Cognitive Psychology*, 3:147–176, 1991.
- [251] T. Valentine, T. Bremen, and S. Bredart. *The cognitive psychology of proper names: On the importance of being Ernest*. London: Routledge. London: Routledge, 1996.
- [252] M. Vignolo. The ontology of products. *Metaphysica*, 11:1–16, 2010.
- [253] M. Vitkovitch, A. Potton, C. Bakogianni, and L. Kinch. Will julia roberts harm nicole kidman? semantic priming effects during face naming. *Quarterly Journal of Experimental Psychology*, 59:1134–1152, 2006.
- [254] M. Vladeanu, M. Lewis, and H. Ellis. Associative priming in faces: Semantic relatedness or simple co-occurrence? *Memory and Cognition*, 34:1091–1101, 2006.
- [255] G. A. V. D. Walle, S. Carey, and M. Prevor. Bases for object individuation in infancy: Evidence from manual search. *Journal of Cognition and Development*, 1:249–280, 2000.
- [256] E. Warrington and T. Shallice. Category specific semantic impairments. *Brain*, 107:829–853, 1984.
- [257] H. Wiese and S. R. Schweinberger. Event-related potentials indicate different processes to mediate categorical and associative priming in person recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:1246–1263, 2008.
- [258] D. Wiggins. *Sameness and Substance*. Oxford: Blackwell, 1980.
- [259] D. Wiggins. Sortal concepts: A reply to xu. *Mind /& Language*, 12:413–421, 1997.
- [260] J. Williams. Is automatic priming semantic? *European Journal of Cognitive Psychology*, 22:139–151, 1996.
- [261] I. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann, 2000.

- [262] J. M. Wolfe and S. C. Bennett. Preattentive object files: shapeless bundles of basic features. *Vision Research*, 37:25–43, 1997.
- [263] D. Wolfram. Term co-occurrence in internet. search engine queries: an analysis of the excite data set. *Canadian Journal of Information and Library Science*, 24:12–33, 1999.
- [264] Y. Xie and D. O’Hallaron. Locality in search engine queries and its implications for caching. INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2002.
- [265] F. Xu. From lot’s wife to a pillar of salt: Evidence that physical object is a sortal concept. *Mind & Language*, 12:365–392, 1997.
- [266] F. Xu. The role of language in acquiring kind concepts in infancy. *Cognition*, 85:223–250, 2002.
- [267] F. Xu. Labeling guides object individuation in 12-monthold infants. *Psychological Science*, 16:372–377, 2005.
- [268] F. Xu. Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences*, 11:401–406, 2007.
- [269] F. Xu and S. Carey. Infants’ metaphysics: The case of numerical identity. *Cognitive Psychology*, 30:111–153, 1996.
- [270] R. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 81:141–145, 1969.
- [271] A. Young, B. Flude, D. Hellowell, and A. Ellis. The nature of semantic priming effects in the recognition of familiar people. *British Journal of Psychology*, 85:393–411, 1994.
- [272] A. Young, D. Hay, and A. Ellis. The faces that launched a thousand slips: everyday difficulties and errors in recognising people. *British Journal of Psychology*, 76:495–523, 1985.
- [273] A. W. Young, A. W. Ellis, and B. M. Flude. Accessing stored information about familiar people. *Psychological Research*, 50:111–115, 1988.
- [274] G. Yovel and N. Kanwisher. The neural basis of the behavioral face-inversion effect. *Current Biology*, 15:2256–2262, 2005.
- [275] H. Zhaoa and S. Ram. Entity identification for heterogeneous database integration - a multiple classifier system approach and empirical evaluation. *Information Systems*, 30:119–132, 2005.