UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Understanding and Exploiting Language Diversity

Khuyagbaatar Batsuren

Advisor

Prof. Fausto Giunchiglia

Università degli Studi di Trento

November 27, 2018

Abstract

Languages are well known to be diverse on all structural levels, from the smallest (phonemic) to the broadest (pragmatic). We propose a set of formal, quantitative measures for the language diversity of linguistic phenomena, the resource incompleteness, and resource incorrectness. We apply all these measures to lexical semantics where we show how evidence of a high degree of universality within a given language set can be used to extend lexico-semantic resources in a precise, diversity-aware manner. We demonstrate our approach on several case studies: First is on polysemes and homographs among cases of lexical ambiguity. Contrarily to past research that focused solely on exploiting systematic polysemy, the notion of universality provides us with an automated method also capable of predicting irregular polysemes. Second is to automatically identify cognates from the existing lexical resource across different orthographies of genetically unrelated languages. Contrarily to past research that focused on detecting cognates from 225 concepts of Swadesh list, we captured 2.7 million cognates across 40 different orthographies and 335 languages by exploiting the existing wordnet-like lexical resources.

Keywords

[Language Diversity, Computational Lexical Semantics, Lexico-semantic Resource, Language Diversity Measure]

Acknowledgements

Thanks to the professors and advisors (Fausto, and Gabor), families (Batsuren Togooch, Erdenedawaa Bal, Temulen Khishigsuren), members(Subashis, Mercedes, Enrico, Hanyu, Sajan, ...), former colleagues (Tsendsuren Munkhdalai, Meijing Li, Park Minghao, Keun Ho Ryu) etc..

thanks to ESSENCE, and other supporting fundings, and contributers (Yue Qin,...) who helped to evaluate and validate the results. The work is supported by European Union ' s 7th Framework Programme project ESSENCE Training Network (GA no. 607062)

Publications

- Giunchiglia, Fausto, **Khuyagbaatar Batsuren**, and Abed Alhakim Freihat. ”One World-Seven Thousand Languages.” In Proceedings of 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-19), 2018. **The best paper award**

- Giunchiglia, Fausto, **Khuyagbaatar Batsuren**, and Gabor Bella. ”Understanding and exploiting language diversity.” In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 4009-4017. 2017.

- Giunchiglia, Fausto, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, and **Khuyagbaatar Batsuren**. ”Crowdsourcing a large scale multilingual lexico-semantic resource.” (2015).

- Munkhdalai, Tsendsuren, Meijing Li, **Khuyagbaatar Batsuren**, Hyeon Ah Park, Nak Hyeon Choi, and Keun Ho Ryu. ”Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations.” Journal of cheminformatics 7, no. 1 (2015): S9.

- Munkhdalai, Tsendsuren, Meijing Li, **Khuyagbaatar Batsuren**, and Keun Ho Ryu. ”Towards a Unified Named Entity Recognition System.” In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 3, pp. 251-255. SCITEPRESS-Science and Technology Publications, Lda, 2015.

- Li Meijing,Tsendsuren Munkhdalai, **Khuyagbaatar Batsuren** and Keun Ho Ryu. A bio-document clustering system based on multiple similarity calculation method. In Proceedings of the 7th International Conference on the Frontiers of Information Technology, Application and Tools, 2014

- **Khuyagbaatar Batsuren**, Tsendsuren Munkhdalai, Meijing Li and Keun Ho Ryu. Keyword extraction using anti-pattern. In Proceedings of the 7th International Conference on the Frontiers of Information Technology, Application and Tools, 2014. **The best paper award**

- Tsendsuren Munkhdalai, Meijing Li, **Khuyagbaatar Batsuren** and Keun Ho Ryu. BANNER-CHEMDNER: Incorporating domain knowledge in chemical and drug named entity recognition. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop vol. 2, 2013

- **Khuyagbaatar Batsuren**, Tsendsuren Munkhdalai, Li Meijing, Namsrai Erdenetuya and Keun Ho Ryu. A novel method for SMS spam filtering using multiple features. In Proceedings of the 6th International Conference on the Frontiers of Information Technology, Application and Tools, 2013. **The best paper award**

- Tsendsuren Munkhdalai, Meijing Li, **Khuyagbaatar Batsuren** and Keun Ho Ryu. A computational approach for biomedical event extraction. In Proceedings of the 6th International Conference on the Frontiers of Information Technology, Application and Tools, 2013

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of language diversity is very well known in the field of historical linguistics and has been studied for many years. Language diversity appears at many levels. Thus, on the level of phonology, while the use of consonants and vowels is a universal feature, the number and typology of these vary greatly across languages [Evans and Levinson, 2009], e.g., from the three vowels of some Arabic dialects to the 10–20 vowels of the English dialects. In morphology, at one end of the spectrum one finds *analytic* languages with very little to no intra-word grammatical structure, such as Chinese. In contrast, *polysynthetic* languages, e.g., some Native American languages [Evans and Sasse, 2002], have sentence-words that other languages would express through phrases or sentences [Crystal, 2004]. On the level of syntax, the various possible orderings of subject, verb, and object have been one of the earliest criteria in linguistic typology. Yet, it was shown that not even these three basic categories are truly universal [Aronoff and Rees-Miller, 2003].

This work has produced a large amount of relevant results with, however, limited practical usability, at least from an Artificial Intelligence (AI) perspective. There are at least two reasons why this has been the case. The first is that, even when using statistical methods, a work in this direction has traditionally relied on low quantities of sample data, one main motivation being the difficulty

of producing high quality large scale language resources. Large scale resources will always be very diversified across languages, more or less complete, more or less correct, more or less dependent on the subjective judgements and culture of the developers. The second is that this work has mainly focused on the syntactic aspects of diversity with much less attention on (lexical) semantics. Exemplar of the state of the art is the recent work in [Youn et al., 2016] which provides a quantitative method for extracting the universal structure of lexical semantics via an analysis of the polysemy of words. The study has been conducted on 22 basic concepts of Swadesh list [Swadesh, 1971] in 81 languages.

At the same time, with the Web becoming global, the issue of understanding the impact of diversity on (lexical) semantics has become of paramount importance (see, e.g., the work on cross-lingual data integration [Bella et al., 2017] and the development of the large multilingual lexical resource *BabelNet* [Navigli and Ponzetto, 2010]). The successes in this area are undeniable, with still various unsolved issues. Thus, for instance, the *Ethnologue* project[1], as of 2017, lists 7,097 registered languages while, to consider the most complete example, as from [Navigli and Ponzetto, 2010], *BabelNet* contains 271 languages. In this respect, it is worthwhile noticing that the languages of the so called *WEIRD* (Western, Educated, Industrial, Rich, Democratic) societies, namely most of the languages with better quality and more developed lexical resources, cannot in any way be taken as paradigmatic of the world languages [Henrich et al., 2010], while many of the not so common *minority languages*, are disappearing from the Web with obvious long term consequences [Young, 2015].

1.1 The Context

The diversity of languages has fascinated researchers and laymen for centuries. More recently, the digital connectedness of the world has brought on needs of

[1] http://www.ethnologue.com

cross-lingual interoperability (in machine–machine, human–machine, and human–human configurations) that have been largely addressed from an AI perspective.[2] At the same time, despite efforts for building multilingual systems, the divide between linguistic *haves* and *have nots*—dominant and minority languages and peoples speaking them—has continued to increase in terms of on-line representation and computational support. This trend is even accelerated by recent massively data-driven approaches that, once again, favour only those languages that can satiate their appetite for gigabytes of digital resources.

1.2 The Problem

To our knowledge, the notion of language diversity has so far been addressed by scientists in two fundamental ways: from a theoretical perspective, historical and comparative linguistics have tried to retrace the genealogical relatedness of languages. Even when using statistical methods, such research has traditionally relied on low quantities of sample data (lexical entries, parse trees, etc.) for the sake of ensuring its very high and controlled level of cleanliness, without which evidence of genealogical relatedness could not be separated from biasing effects of culture or environment. From an applied perspective, AI and more specifically computational linguistics have dealt with language diversity through a one-by-one effort, by adapting generic methods to individual languages (e.g., machine learning sequence labellers and parsers, lexicons, wordnets, or lately word embeddings). While the success of this approach has been undeniable for dominant languages, the cost of cross-lingual adaptation is generally very high and cannot be paid by all linguistic communities.

Specific subproblems of cross-lingual interoperability have been successfully addressed by automated efforts such as statistical machine translation (*Google Translate*) or the large multilingual lexical database *BabelNet* [Navigli and Ponzetto,

[2]From the early days, machine translation has been an emblematic problem for AI research.

2010]. However, because of the fully automated and one-size-fits-all approach such resources tend to suffer from bias towards knowledge embedded in Western (mainly Anglo-Saxon) language and culture and from a general lack of exploitable data for languages with lesser online presence [Vossen et al., 1999].

1.3 The Solution

Our research aims to provide the missing link between these two major viewpoints. We demonstrate that modern large-scale linguistic resources can be leveraged for the purposes of the *analysis* of linguistic diversity and, in turn, results of such analyses can help us overcome problems of incorrectness and incompleteness of resources by *synthesising* new, 'diversity-aware' knowledge with a precise understanding of the range of languages to which it applies.

The core of our methodology is a formal, quantitative *measure of diversity* of a set of languages. There are, of course, several manners one could define the notion of language diversity: culturally, geographically, based on their genealogical relatedness, taking a synchronic or a diachronic perspective. We consider diversity as the combination several (cultural, geographic, etc.) factors and take a compositional approach to defining it; in this thesis we present a first approximation that we plan to refine in further research. The diversity measure can be applied as a general device to *experiments* on various linguistic phenomena, concluding on their universality or, on the contrary, locality to a subset of languages. This serves as evidence for the scope of applicability of automated methods that address problems of incompleteness or incoherence in existing resources.

Human validation of automatically obtained results is, of course, crucial and we consider it as an integral part of each experiment. For certain use cases we view crowdsourcing as the ideal long-term solution for ensuring coherence with the common-sense perception of language, in congruity with our commitment to

a research programme motivated by real-world applications [Giunchiglia et al., 2015].

In this thesis, among the many structural levels of language we turn our attention to lexical semantics: the diversity of word meanings across languages. In particular, we examine the universality of *semantic relatedness* between concepts. Our motivation is to use these results to extend lexico-semantic resources, such as the very widely used language-specific *wordnets*, and also to clean them from some of the noise introduced by previous automated approaches. In this way we address the two most common problems of multilingual lexico-semantic resources: incompleteness and linguistic bias.

1.4 Outline of the thesis

The rest of this thesis is organised as follows:

Chapter 2 represents the state-of-the-art work in different research fields that are related to this thesis work.

Chapter 3 describes the Universal Knowledege Core (UKC) that is a multilingual lexical resource, used in the experiments of our studies.

Chapter 4 describes a set of quantitative, formal measures of language diversity in different levels of attributes, namely: genetic diversity and geographic diversity.

Chapter 5 represents a set of quantitative, formal measures of incompleteness and incorrectness of lexical resources. All those quantitative measures are being used in the next chapters of case studies.

Chapter 6 represents a longstanding problem of "polysemy vs homonym", and how the quantitative measures are efficient to treat this problem.

Chapter 7 shows how more lexical relations could be discovered precisely by exploiting PWN.

Chapter 8 represents the identification of cognates is also a longstanding

problem in historical and comparative linguistics, and how the quantitative measures are efficient to treat this problem. Chapter 9 concludes this thesis.

Chapter 2

State of the Art

This PhD thesis work has a broad interdisciplinary research area. In this chapter, we presents the state-of-the arts in three main areas, namely: Multilingual lexical and semantic knowledge resources (Chapter 2.1), Universality of Languages in Lexical Semantics (Chapter 2.2), and Comparative Linguistics (Chapter 2.3).

2.1 Lexical Resources

Large-scale multilingual lexical resource is crucial for us to investigate whether the proposed hypotheses and approaches are efficient and useful for overcoming the issues of linguistic phenomena. One of the famous lexical resources developed first is Princeton WordNet (PWN) [Miller et al. 1993] in English. One of key secrets in its development is based on the psycholinguistic theories as a result of Miller's forty year research in the psycholinguistic field. PWN has been proved to be very robust and efficiently useful for many NLP applications. This success attracted many researchers and professors in other countries to develop wordnets in their native languages. The linguistic resources are categorized into two kinds: (a) monolingual resources and (b) cross-lingual resources.

2.1.1 Monolingual lexical resources

For each language studied, these should provide us with a lexicon (as large as possible) as well as lexico-semantic relations across lexical entries such as synonymy, polysemy, derivational relatedness, etc. Well-known and widely available such wordnets are [Isahara et al. 2008] in Japanese, [Lindén and Niemi, 2014, Lindén et al., 2012] in Finnish, [Huang et al., 2010] in Chinese, [Black et al., 2006] in Arabic, [Koeva et al., 2004] in Bulgaria, [Pociello et al., 2011] in Basque, [Pedersen et al., 2009] in Danish, [Garabík and Pileckytė, 2013] in Lithuanian, [Postma et al., 2016] in Dutch, and many more. Although we found 40 monolingual wordnet resources from Internet, a majority of the resources have a very low coverage in number of word, not greater than 5000 words. This incompleteness limitation, comparing with PWN, makes even harder to use those resources for multilingual semantic applications, and explains why NLP applications with those resources for the minority languages have poor performances.

2.1.2 Multilingual lexical resources

These provide a fine-grained mapping between lexemes in different languages. Some projects have developed the multilingual lexical resources manually that include namely: MultiWordNet [Pianta et al., 2002b], EuroWordNet [Vossen, 1998], Multilingual Central Repository [Gonzalez-Agirre et al., 2012], BalkaNet [Tufis et al., 2004], IndoWordNet [Bhattacharyya, 2017] and others. In addition, several projects, e.g. Open Mutliingual WordNet [Bond and Foster, 2013], have combined all existing wordnet-like resources into one lexico-semantic database while many others [Matuschek et al., 2013] have merged them with some richer knowledge resources including Wikipedia, Wiktionary, and OmegaWiki. In the following, I would like to focus on the two main state-of-the-art communities for multilingual lexical resources.

First. A few years ago, Global WordNet Association (GWA) is established as a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world. GWA organizes the global wordnet conference biannually to bring together researchers to discuss the emerging issues around wordnets. The organization put a lot efforts to encourage researchers to make their private wordnets open and publicly available, and more importantly free to use for any purpose of research. In recent years, the community designated the Global WordNet Grid [Vossen et al., 2016, Bond et al., 2016] to capture a diversity of semantics of all languages on wordnets.

Second. BabelNet [Navigli and Ponzetto, 2010] is the largest and famous one on a type of multilingual lexico-semantic resources, obtained from the automatic integration of several language resources, namely: WordNet, Open Multilingual Wordnet, Wikipedia, and OmegaWiki. Currently, this resource covers 284 languages, 6 million concepts and 785 million words. Its success on the semantic space of concepts is undeniably a huge contributions to many applications using this resource. However, a linguistic space of the resource is currently limited to the morphological relations of English as same as other multilingual resource s like OMW and IndoWordNet.

In this thesis, one of our main purposes is to enrich the linguistic space of the existing lexical resources. We focus to enrich the following linguistic dimensions:

- Polysemy vs Homonym: is described in chapter 6
- Morphological relations: is described in chapter 7
- Cognates: is described in chapter 8

2.2 Language Diversity in Lexical Semantics

This field of lexical semantics is dealing with how concepts and meanings are expressed by words. Interestingly, to express new meaning in a natural language, the speech communities have no need to invent new words, but can extend existing words beyond its core meanings by adding morphological patterns (morphosemantic) [Fellbaum et al., 2007], using same word (polysemy) [Srinivasan and Rabagliati, 2015] or multiwords (e.g. computer mouse) [Hsieh and Chang, 2014].

From this point of view, the verbalization process of new meaning is sometimes systematically related to other meaning that relates to the meaning semantically. In this thesis, we want to investigate that a semantic relationship exists between two concepts if they share such properties in unrelated languages as diverse as possible, but our hypothesis have been questioned by whether such results exist across languages due to universal properties of human cognition, as opposed to the particulars of cultural history or local environments.

The universality of linguistic phenomena has been in the focus of historical and comparative linguistics, as well as of the related field of linguistic typology [Croft, 2002]. In this context, proper language sampling was crucial to avoid biased results (as explained in chapter 4), hence the development of quantitative measures of diversity as in [Bell, 1978, Rijkhoff et al., 1993] that also inspired our work. Measures of geographic, climatic, and cultural relatedness were used in [Youn et al., 2016] in a somewhat more sophisticated manner than our embryonic geographic diversity measure. Universality has been most famously researched on the syntactic level in search of a *universal grammar* [Evans and Levinson, 2009] but also in the lexicon. Classic quantitative approaches as described in [McMahon and McMahon, 2005], such as lexicostatistics [Swadesh, 1955], mass comparison [Greenberg, 1966], or the recent paper [Youn et al., 2016] on the universality of semantic networks, perform comparisons on rela-

tively small (of up to a couple hundred entries) but very carefully selected word lists expressing the same meaning across a large and unbiased language sample (e.g., the *Swadesh list* [Swadesh, 1971]). Our research, on the contrary, takes the results of experts on genetic relationships as granted for our diversity measures. Beyond understanding the diversity of the language sets we are working on—and thus evaluating the scope of cross-lingual applicability of our results—we have no a priori reason to exclude certain types of words or phenomena from our experiments and can leverage entire lexicons available to us. *The intuition is that the scale of the resource will average out local biases.*

The study of polysemy also has a long history, see, e.g., [Apresjan, 1974, Lyons, 1977]. In particular, various computational methods have been proposed for the prediction and generation of polysemy instances from regular (productive) patterns [Buitelaar, 1998, Peters, 2003, Srinivasan and Rabagliati, 2015, Freihat et al., 2016]. Our study goes beyond the limitation of regularity as our goal is not to create rules to be applied over classes of concepts but, rather *to find widely recurring polysemy patterns across multiple languages* with respect to specific concept pairs.

2.3 Databases for Comparative Linguistics

In recent years, many data resources for comparative linguistics have been released in the machine readable formats. In the following subsections, we provide a sample of the most popular databases in this field that are publicly available and connected to our study.

2.3.1 Expert language classifications

At this time, the most comprehensive databases of languages are only two, namely: Ethnologue [Simons, 2017] and Glottolog [Hammarström et al., 2015].

- *Ethnologue* is a web-based genetic classification database of 7,097 languages. It also contains many supporting informations of each language, including countries, dialects, a number of speakers, linguistic affiliations, and locations. Although this database has extremely good quality, only a small part of the information is publicly available in a digital form.
- *Glottolog* is a genetic classification of 7,943 languages and dialects, that are evidently linked to the bibliographic references of about 180,000 linguistic studies such as grammars, dictionaries, word lists, texts etc.
- *WALS* is a genetic classification of 2,679 languages and dialects, alongside with the supporting linguistic informations (phonological, grammatical, lexical), and geographic locations. The linguistic information of all languages are evidently linked to the bibliographic references of about a team of 55 authors.

2.3.2 Cognate databases and word lists

In this subsection, we review the popular cognate databases and word lists that are often used for automatic cognate identification methods.

- Automatic Similarity Adjustment Program (*ASJP*[1]) database [Wichmann et al., 2010] is a collection of wordlist of 7655 languages and dialects for 40 basic concepts (e.g. sun, person, you). In overall, it contains 307,396 words that are given in form of phonetic transcription, but the words in original orthography are missing, and this restriction limits data integration with lexical resources.
- *Indo-European Lexical Cognancy Database*[2] collected wordlists of 163 Indo-European languages for 225 basic concepts of Swadesh list. Word entries are given in both forms of orthography and IPA transcriptions, and also

[1] https://asjp.clld.org/
[2] IELex; http://ielex.mpi.nl

words are manually assigned cognate classes. Even though Indo-European phylum is orthographically rich , the database supports a very few of orthographies including latin, cyrillic, and greek but not to arabic, hindi, urdu, odia and other indo orthographies.

- *Austronesian Basic Vocabulary Database* collected wordlists of 1467 pacific-region languages for 210 basic concepts of Swadesh list. A majority of the 1467 languages belongs to Austronesian phylum. Word entries are given in phonetic transcriptions, and also words are manually assigned cognate classes.

In this study of the thesis, we integrated all the publicly available data of Ethnologue, WALS, and Glottolog databases, and want to contribute a new cognate database to the existing body of the cognate wordlists mentioned above. This new database is extracted from the existing wordnet-like lexical resources, and the details are given in Chapter 8.

Chapter 3

Universal Knowledge Core

Our first goal in this thesis is to describe a multilingual lexical resource that we call the *Universal Knowledge Core (UKC)*.[1] We have used this resource to conduct this experiments and test the method proposed in the following chapters. The UKC shares all the PWN design choices but one: the *synsets* which in different languages codify the same meaning are clustered into *language agnostic concepts*. Furthermore, in the UKC, semantic relations link concepts, and not synsets, and create a *language independent semantic network*, that we call the *Concept Core (CC)*. So far, the UKC has evolved as a combination of importing of freely available resources, e.g., WordNets or dictionaries of high quality, and language development, see e.g., [Giunchiglia et al., 2017]. As of to day, it contains 335 languages, 1,333,869 words, 2,066,843 synsets and more than 120,000 concepts. Table 1 reports the distribution of words over languages where, more or less, 90% of the words belong to 50 languages.[2]

The existence of the CC makes the UKC *not biased by any language and culture* and, therefore, *inherently open* and easily extensible. For instance, *lexical gaps*, namely previously missing concepts lexicalized in a new language can be dealt with by adding a new concept, thus solving one of the difficulties

[1] The word *knowledge* in *UKC* is motivated by our focus on studying language not *per se* but as a key component of reasoning systems.

[2] From February 2018, the UKC will be browsable on line at the link http://kidf.eu.

Table 3.1: Language Distribution.

#Words	#Languages	Samples
>90000	2	English, Finnish
>75000	4	Mandarin, Japanese, etc.
>50000	6	Thai, Polish, etc.
>25000	17	Portuguese, Slovak, etc.
>10000	29	Islandic, Arabic, etc.
>5000	39	Swedish, Korean, etc.
>1000	66	Hindi, Vietnam, etc.
>500	85	Kazakh, Mongolian, etc.
>0	335	Ewe, Abkhaz, etc.

which arise in the construction of multilingual Wordnets. This is crucial given that the languages of the so called WEIRD (Western, Educated, Industrial, Rich, Democratic) cultures cannot in any way be taken as paradigmatic of the world languages [Henrich et al., 2010]. Furthermore, it is also important to notice how the co-existence of synsets and concepts allows for the seamless integration of language dependent and language independent reasoning. Thus, on one side, any application using concepts will automatically run for any language supported by the UKC, see, e.g., the work on cross-lingual data integration described in [Bella et al., 2017], while, on the other side, as discussed in detail in Section 3, synsets can be used to keep track of the local language and culture. An exemplary application is the extension to multiple languages fof the work in [Deng et al., 2009, 2014] which uses Wordnet for the large scale classification of photos (what is depicted by a photo is biased by culture; compare, e.g., the photo of a home in Italy with that of a home in Mongolia).

3.1 Language and Concept Core

The key design principle underlying the UKC is to maintain a clear distinction between the *language(s)* used to describe *the world as it is perceived* and what

is being described, i.e., the world itself. The *Concept Core (CC)* is the UKC representation of the world and it consists of a semantic network where the nodes are language independent *concepts*. Each concept is characterized by a unique identifier which distinguishes it from any other concept. The semantic network consists of a set of semantic relations between nodes which relate the meanings of concepts, where these relations are an extension of those used by the PWN (e.g., *hyponym*, *meronym*).

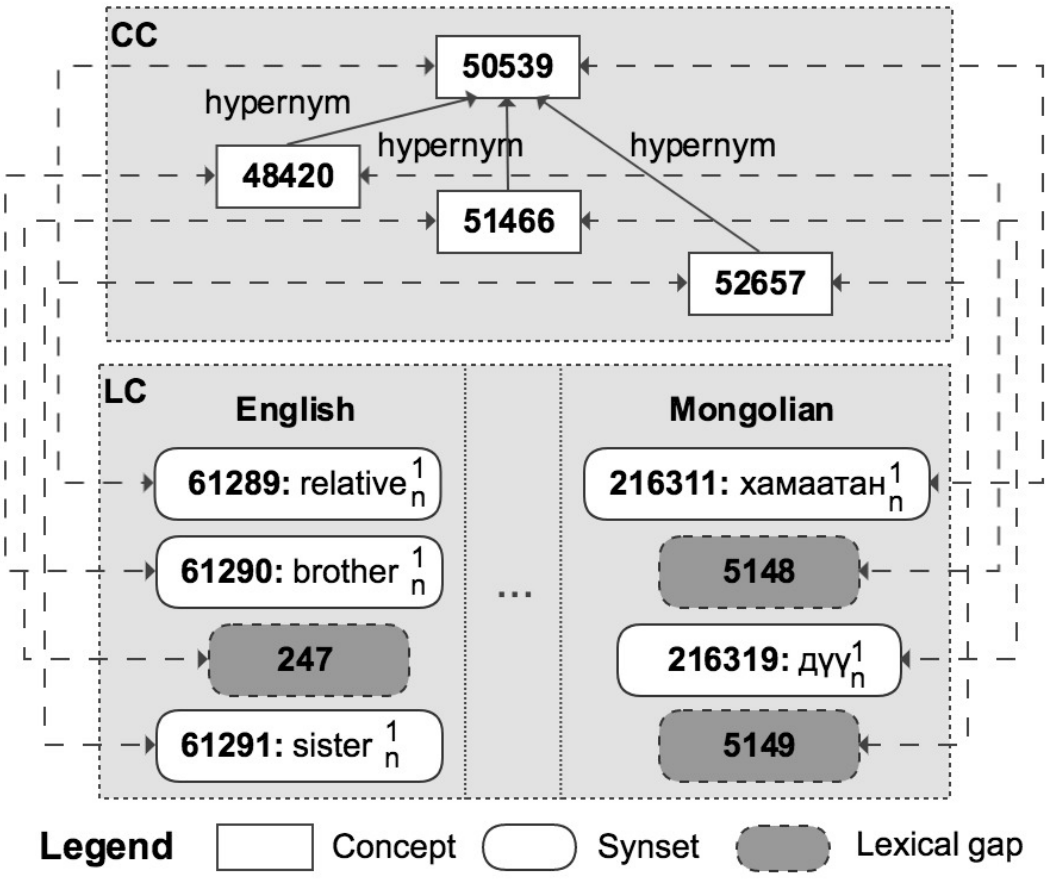

Figure 3.1: A fragment of the semantic network of concepts and their synsets.

We talk of the *Language Core (LC)*, meaning the component that, in the UKC, corresponds to the PWN, namely the set of *words*, *senses*, *synsets*, *glosses* and *examples* supported by the UKC. Despite playing a similar role, the LC is actually quite different from the PWN. Similarly to the PWN, in the LC each synset

is univocally associated with one language and, within that language, with at least one word. Differently from the PWN, synsets are linked to concepts, and there is the constraint that each synset is linked to one and only one concept. There is, furthermore, the constraint that, *for a concept to be created, there must be at least one language where it is lexicalized*. Given the multilinguality of the UKC, there is a one-to-many relation between concepts and synsets. Figure 1 shows how synsets and concepts are related ("n" means that the reference word is a noun, "1" that that synset is associated to its first sense).

Glosses and *examples* are associated with synsets, as in the PWN. We have evaluated the possibility of associating glosses also to concepts. Ultimately, we decided that this should not be the case as such a description would be linguistic in nature and there is no universal language which could be used to describe all the concepts in the CC. One difference with the PWN is that, in the UKC, lexical gaps have glosses, even if they do not have examples (which would be impossible). The intuition is that the gloss of a lexical gap can be seen as "local" language dependent description of a missing synset. This choice has turned out to be pragmatically useful when one is interested in understanding the meaning of a lexical gap without knowing the language(s) which generate(s) them.

3.2 Words, Synsets, and Concepts

Humans build representations of what they perceive, what we usually call *the world*, as complex combinations of *concepts* where, following [Giunchiglia and Fumagalli, 2016], we take *concepts* to be *mental representations of what is perceived*. The recognition of a concept is taken to be the result of (multiple) *encounters*, i.e., events, $e_1, ..., e_n$, during which *substances manifest* themselves to a perceiver (e.g., an observer or a listener), where substances have two fundamental properties:

1. *they maintain some level of, but not full, invariance on how they manifest*

themselves to observers across multiple encounters and

2. *this ability is an intrinsic property of substances*.

Examples of concepts generated from substances are objects (e.g., persons, cars, cats), actions (e.g., walk, drive) roles (e.g., father, president); see [Giunchiglia and Fumagalli, 2016, 2017] for a detailed discussion about these notions and also [Millikan, 2000] for the early work in the field of *Biosemantics* which introduced the notions of substance and encounter. The *key* observation is that we take concepts as representations denoting *sets of encounters*, rather than *sets of instances* which share a set of properties, as it is the case in the *Descriptionistic* theories of meaning, e.g., *Knowledge Representation* or the "usual" *logical semantics*. Thus, for instance, the denotation of the concept *car* is the set of times a car has been perceived, e.g., seen by me, rather than the set of cars which, e.g., are in Trento. This shift allows us to treat concepts and words uniformly. We take words, like concepts, to be representations of the world; more specifically, to be mental representations of mental representations of the world (i.e., of *concepts*). As such, words, like concepts, are the results of sets of encounters $e_1, ..., e_n$ during which they are perceived by, e.g., a listener of reader, as produced by, e.g., a human speaker or written text. Thus, for instance, analogously to what happens for the *concept car*, the *word car* denotes the set of input occurrences which are generated by looking at a set of documents and/or by hearing a set of utterances.

We represent words, synsets and concepts and their respective roles as in Figure 3.2. Outside the UKC there is the world as we perceive it, e.g., via vision (bottom) or listening (top). At the bottom there are concepts $c_1, ..., c_n$, while, at the top, there are words $w_1, ..., w_n$ (in Figure 2, *car* and *automobile*), where both words and concepts are perceived as the result of the encounters $e_1, ..., e_n$.

Moving to the center of Figure 2, the synsets $s_1, ..., s_n$ are linked to words and to concepts, see, e.g., the word *car* in Figure 2. We call these two links *word sense* and *concept sense*, respectively, or simply *sense*, when the context

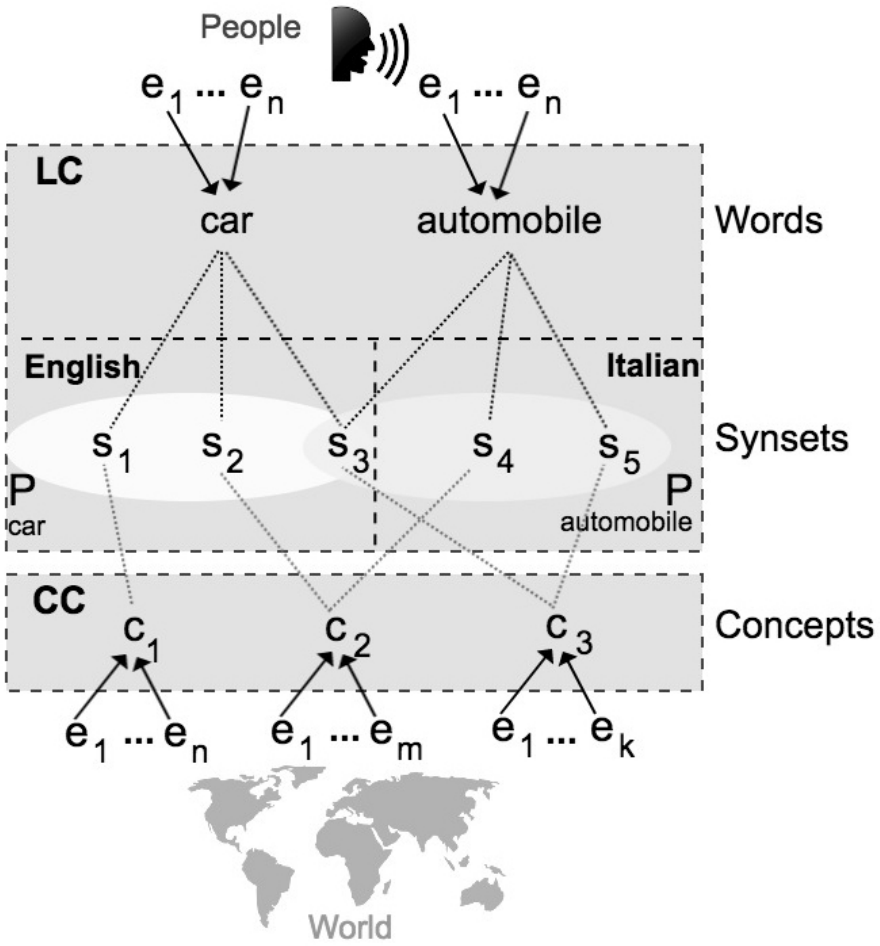

Figure 3.2: The UKC and the World.

makes clear what we mean. Notice how, as represented in Figure 2, both words and concepts are ambiguous representations of synsets, in the sense that there is a one-to-many relation between them and synsets. The sense of a word depends on the *context* within which it is perceived while the sense of a concept depends on the *language* used. Thus, as in Figure 2, the word *car* and the word *automobile* denote the sets of synsets P_{car} and $P_{automobile}$, respectively, where each synset is indexed by a different context, and these two sets overlap in s_3. In turn, the concept c_3, like any other concept, is denoted by a set of synsets, each synset belonging to a different language (English and Italian in Figure 2). c_1, being non being lexicalized in Italian, is a probe for a possible lexical gap in this

language. Notice also how there are words, e.g., *automobile* which are shared across languages, this being pervasive with languages with common roots, e.g., Portoguese and Brasilian Portoguese.

As a result the UKC implements the following stratified theory of meaning:

- the results of perception, i.e., words and concepts, denote the set of events during which they are perceived; they define the boundary between the UKC and the world;
- words denote sets of synsets, one per context of use;
- synsets denote concepts, where any concept is denoted by multiple synsets, one per language;
- Any triple $\langle w_i, s_i; c_i \rangle$, with s_i word sense of w_i and c_i concept sense of s_i, is a *Causal connection* $CC(w_i, c_i)$ between w_i and c_i.

$CC(w_i, c_i)$ implements the *causal connection between words and concepts* that humans exploit in knowledge representation and reasoning. Given that media, e.g., photos and videos, are direct representations of concepts, the above organization paves the way to integrated multimedia and multilanguage systems, extending the work in the integration of linguistic resources and media, so far done only for single languages, see, e.g., [Deng et al., 2009, 2014].

3.3 World, Language(s), and Model(s)

The three-layer organization of meaning into words, synsets and concepts, as represented in Figure 3.2, motivates a three layer design of the UKC, as represented in Figure 3.3, with the first two layers inside the LC and the third inside the CC. We have:

1. the *Word Layer*, which stores what we call the *Universal Lexicon*,

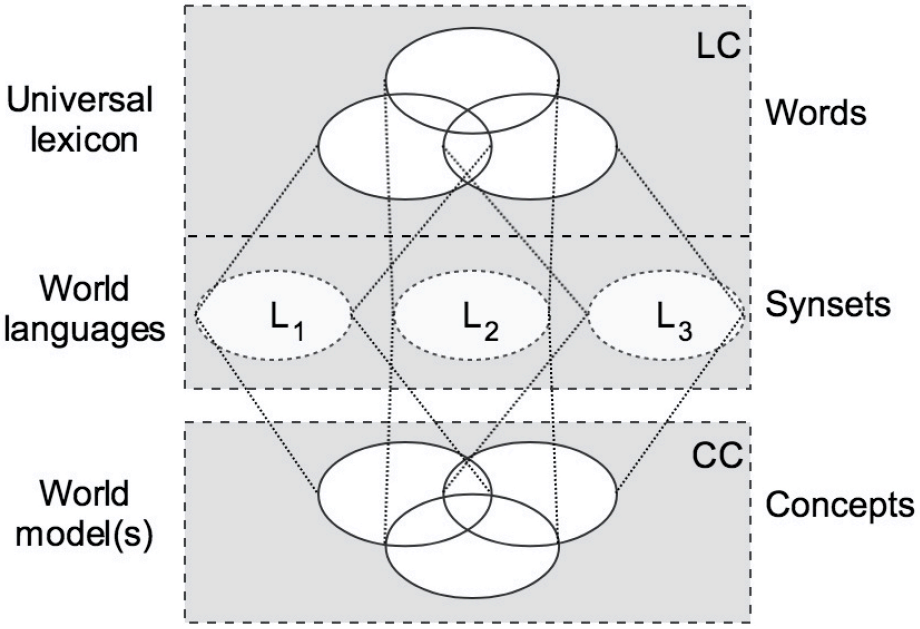

Figure 3.3: Languages, Universal lexicon and World model(s).

2. the *Synset Layer*, which stores the *World Languages*, and

3. the *Concept Layer*, which stores the *World (mental) model(s)*, as represented by the CC.

Word Layer and *Concept Layer* store the results of perception while the *Synset Layer* implements the causal connection between words and concepts. In the *Synset Layer* each circle represents a different language where all languages are mutually disjoint, this being a consequence of the fact that, differently from what is the case for words (see Figure 2), each synset is associated with one and only one language. On this basis, in the UKC, we formally define *a Language as a set of synsets*, in formulas

$$L = \{s_i\}_{i \in I_L}.$$

The above definition is at the basis of all the definitions regarding language diversity and resource quality provided in the next sections. It allows, for instance, to compare the concepts lexicalized in the different languages, including

the absence of lexicalizations (which are probes for lexical gaps) and to study how polisemy and synonymy map to the underlying concept semantic network.

The *Word Layer* stores the *Universal Lexicon*, namely the set of the words belonging to at least one language. Notice that a word, meant as the event by which it is perceived and recognized, does not a priori belong to any language. It is only a *sign* or a *sound* which may be used in more than on language and which is recognized as belonging to a language as part of the word sense disambiguation process. Of course, as represented in Figure 3.3, it is possible to reconstruct the set of words of a Language from synsets using the inverse of the word sense relation.

The *Concept Layer* is a language agnostic representation of the world as we perceive it. *But, a model generated by who?* In the UKC, the world, as we perceive it, is taken to be a source of *perception events*. By perception event we mean here the concrete sensing action, performed by a sensing subject, which generates concepts and words, and the causal relations linking them. This gives us the possibility to define the notion of world (as we perceive it) in terms of the subject(s) which actually perform the sensing actions enabling the perception events.

According to a first mainstream interpretation, the CC is the model of the entire world, as it is generated by all the people (speaking all the languages) in the world. However, according to a second interpretation, the CC can also be seen as the union of the models of the world as they are generated by the different people (speaking the different languages) in the world, as represented in Figure 3.3, e.g., the models of the 7,097 languages registered by the *Ethnologue project*.[3] Clearly these models intersect and are a subset (a subgraph) of the overall CC. It is interesting to notice how this view can be easily pushed to the extreme by associating a different world model to any different sensing subject (e.g., any person). During the generation of a lexicon, lexicographers would

[3] http://www.ethnologue.com/

choose from a "common pot" words and concepts, namely what we all share via perception, while, at the same time, they would be able to decide synsets and senses, namely the causal relation from words to objects that is unique to each of us.

In this perspective, notice how *lexical gaps* are core to our studies on language diversity as they provide evidence of the different worlds perceived by people speaking different languages. The notion of lexical gap is seemingly quite intuitive: a lexical gap is a missing element in the lexicon. *But, missing with respect to what?* The approaches that we are aware of define this notion in terms of properties of the lexicon. Thus for instance, [Kjellmer, 2003] defines lexical gaps as holes in the systematicity of languages while [Bentivogli and Pianta, 2000, Cvilikaitė, 2006] define them as the lack of lexicalization detected when comparing two languages. [Lehrer, 1970] defines a lexical gap as a missing lexicalization of the semantic structure of a language, based on the analysis of the lexicon of that language. Our notion of lexical gap codifies directly the fact that a lexical gap is a missing link between a lexicon and semantics: *a lexical gap is a concept for which a language is known not to have a synset*. Notice how this definition relates directly to how different cultures generate, via language, different world models.

3.4 Lexical gaps

The absence of a certain concept in a language is motivated by two meanings. One is a lack of physical encounters that speakers of a language have very little or no experience to think of that concept. The second is a concept's redundancy that in the language l other concepts take its place of a semantic field.

Definition 1. Lexical gap A pair of concept[4] and language $<c, l>$ is a ***lexical gap*** if and only if a lexicon of language l lacks a word to express a concept c.

[4]Here, a concept is referring to lexical concept, lexicalized at least one language.

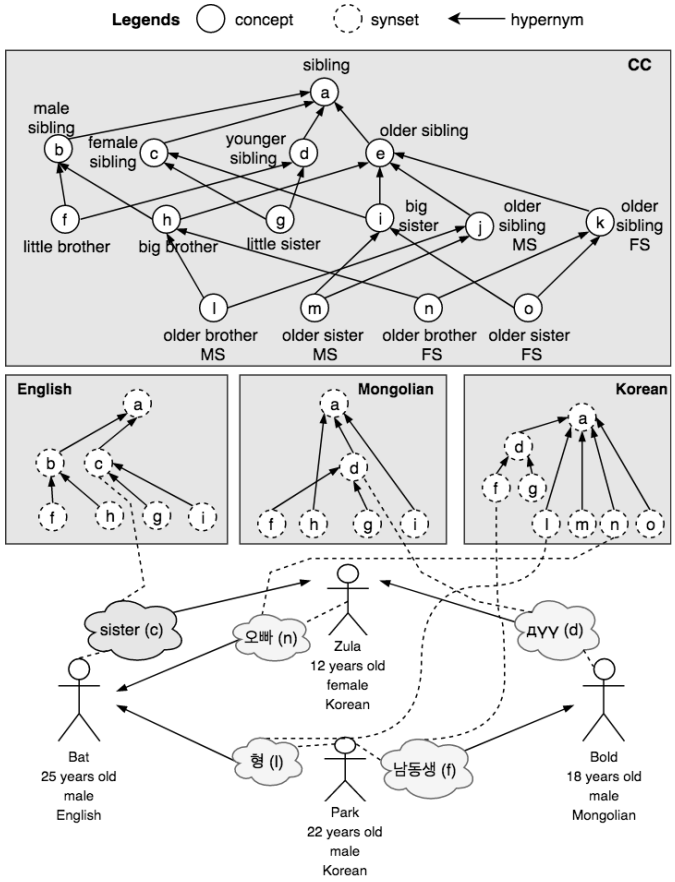

Figure 3.4: Example of four siblings of one family with the semantic field of "sibling". (MS is a male speaker and FS is a female speaker.)

The definitions of lexical gaps are vary from one another [Lehrer, 1970] [Bentivogli and Pianta, 2000]. According to one definition , a lexical gap means that a language expresses a certain concept with a *free combination of words*. By that definition, it excludes many cases of cultural concepts where a language tends to directly borrow a foreign word to express a certain concept instead of coining a word or a free combination of words. For example, let's suppose that a japanese sport 相撲 */sumō/* is being introduced to English. In such case, a language speakers have often naturally borrowed a foreign word to express that meaning.

Definition 2. World gap A lexical gap $g =<c, l>$ is a *world gap* if speakers

of the language l lacks physical encounters on substances of the concept c. For example, a hundred years ago, a chinese martial art kung fu (功夫) were not introduced to many countries of today.

Definition 3. Representation gap A lexical gap g =<c, l> is a *representation gap* if lexical encounters of the language l, representing substances of concept c, are pointing to one or more different concepts than c.

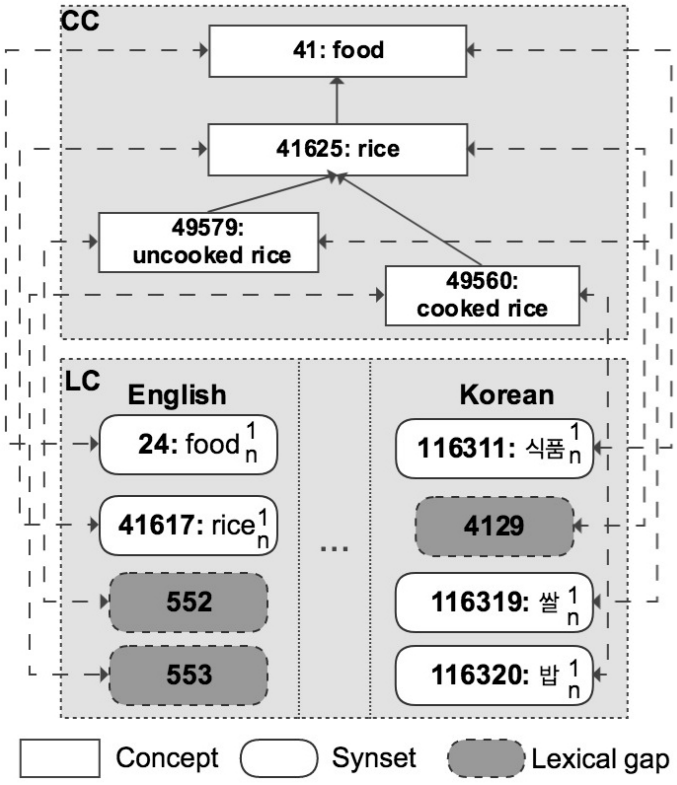

Figure 3.5: Example of hypernym gaps for the concepts of rice.

Definition 4. Hypernym gap A representation gap g =<c, l> is a *hypernym gap* if least one hyponym or specific concept of c is lexicalized in the corresponding language l. For instance, a concept "rice" is a hypernym gap in Korean because in that language its more specific concepts are lexicalized by the words " "(*ssal - uncooked rice*) and " "(*baap - cooked rice*).

Definition 5. Hyponym gap A representation gap g =<c, l> is a *hyponym*

gap if any hyponym or specific concept of c is a lexical gap in the corresponding language l.

Linguistically, a hyponym gap can be recognized as a functional gap by the definition given by Lehrer [Lehrer, 1970], but also includes an attributional gap. The reason why its all hyponym concepts are lexical gaps is that in the lexical inheritance system employed by UKC and PWN, any hyponym concept inherits the distinguishing features (e.g. attributes and functions) of a concept c, so if an inherited attribute or function itself is a gap in language then its hyponym concepts should be a lexical gap in principle. For instance, in Figure 3.5, the concepts *j* and *k* are hyponym gaps in English and Mongolia because the attribute of gender-based speakers is absent in those languages.

Chapter 4

Language Diversity in Lexical Semantics

> "A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language."
> —-*Noam Chomsky*

The problem of quantifying the diversity of languages is not new, see, e.g., [Bell, 1978, Youn et al., 2016]. Our ideas build upon the work described in [Rijkhoff et al., 1993]. The main goal of this work was to construct balanced datasets with the goal of avoiding linguistic bias. Still sharing the same intuitions, we work in the other direction. Namely, we have the data sets and we measure their diversity in order to exploit it in the solution of well-known linguistic problems.

Diversity has many causes. To name some: genetic ancestry (languages with common origins), geography (due to the influence of physical closeness), culture (effects of cultural dominance). In this thesis we present a first attempt at quantifying a global *combined diversity measure* in terms of *genetic diversity* and *geographic diversity*. Given a language set $\mathcal{L}$, we define its combined diversity measure as follows:

$$\text{ComDiv}(\mathcal{L}) = \text{GenDiv}(\mathcal{L}) + \beta\text{GeoDiv}(\mathcal{L}) \tag{4.1}$$

In the equation above $\beta \in [0, 1]$ normalizes the effects of genetic diversity over those of geographic diversity. We compute the *Relative (Combined) Diversity* of two languages by taking $|\mathcal{L}| = 2$ and we (generically) say that *two or more languages are similar when they are not diverse* and we extend this terminology to all forms of diversity. Let us define the notions of genetic and geographic diversity in the following subsections.

4.1 Genetic Diversity

Languages are organized in a *Language Family Tree* which represents how, in time, languages have descended from other languages, starting from the ancestral languages [Bell, 1978]. A fragment of this tree is shown in Fig.4.1. This figure must be read as follows. The root is a placeholder for collecting all languages. Labeled intermediate nodes are sets of languages (phyla or families) where the label is the name of the set. Unlabeled intermediate nodes correspond to missing names of language sets and serve the purpose of keeping the tree balanced (crucial for the computation of diversity, see below). Leaves denote languages. In general, we write $\mathcal{T}(\mathcal{L})$ to mean the family tree $\mathcal{T}$ for the set of languages $\mathcal{L}$ (when clear we drop the argument from $\mathcal{T}$).

4.1.1 Quantifying Genetic Diversity

The idea behind the computation of *genetic diversity* is that languages that split closer to the root (that is, further back in time) will have more fundamental changes than those involved in the more recent splits. We capture this intuition by pondering each node n in the Language Family Tree by a real number that decreases with the distance from the root. Thus languages which split very early will generate multiple long branches, thus increasing the overall diversity value. While [Rijkhoff et al., 1993] used linearly decreasing weights, we have chosen the inverse exponential of $\lambda^{-\text{depth}(n)}$ where the depth of the Root is 0 and thus

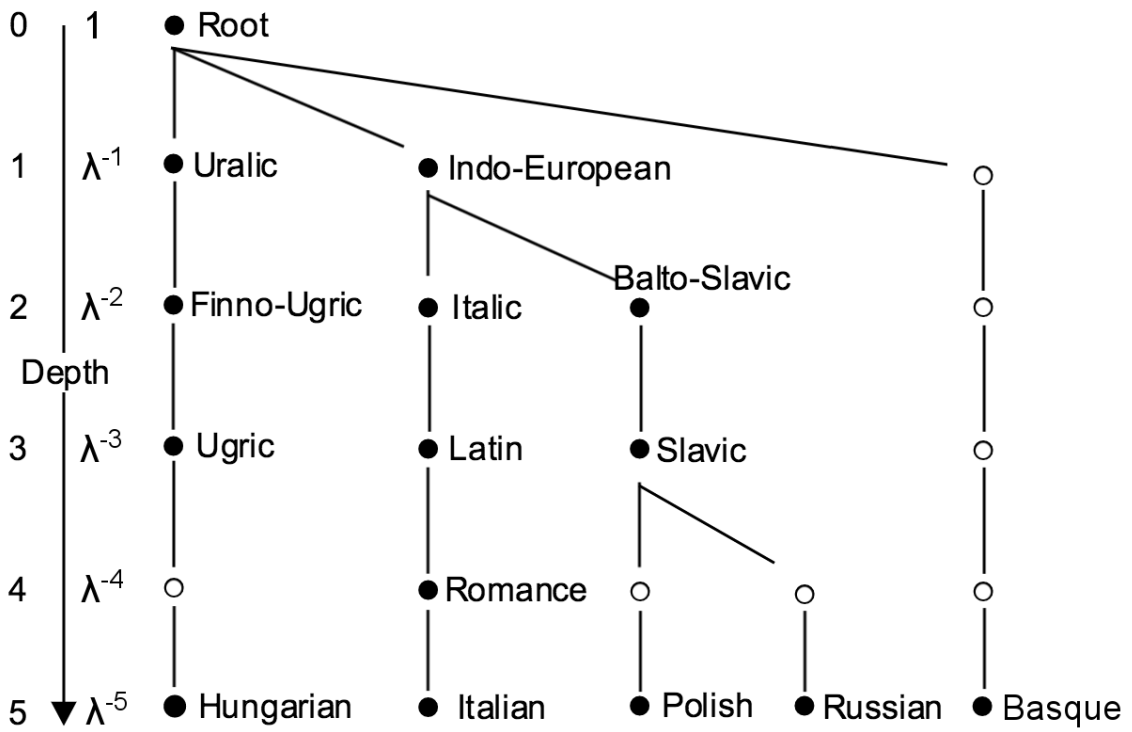

Figure 4.1: A fragment of the phylogenetic tree.

its weight is 1 and where, below it, each phylum is weighted $1/\lambda$, then $1/\lambda^2$, and so on. Furthermore we normalize GenDiv to be in the range [0,1]. More specifically, let $\mathcal{T}(\mathcal{E})$ be the family tree of a reference set of languages $\mathcal{E}$, which in our case we take to be the languages in the UKC. Let $\mathcal{L} \subseteq \mathcal{E}$ be a set of languages for which we want to compute the diversity level and $\mathcal{T}(\mathcal{L})$ the corresponding *minimal subtree* of $\mathcal{E}$. Then, the genetic diversity of $\mathcal{L}$ is taken to be 0 if $|\mathcal{L}| < 2$, and, otherwise, defined as:

$$\text{AbsGenDiv}(\mathcal{L}) = \sum_{n \in \mathcal{T}} \lambda^{-\text{depth}(n)} - 1 \tag{4.2}$$

$$\text{GenDiv}(\mathcal{L}) = \frac{\text{AbsGenDiv}(\mathcal{L})}{\text{AbsGenDiv}(\mathcal{E})} \tag{4.3}$$

where AbsGenDiv is what we call this function the *Absolute Genetic Diversity* and AbsGenDiv($\mathcal{E}$) is the *Reference Genetic Diversity*. To provide some exam-

ples, assume we take $\lambda = 2$. Then AbsGenDiv$(\mathcal{E}) = 88.127$ and GenDiv$(\mathcal{E}) = 1$, while, with $\mathcal{L}_1$ ={Hungarian, Italian, Polish, Russian, Basque} (the languages in Fig. 4.1) we have AbsGenDiv$(\mathcal{L}_1) = 3.469$ and GenDiv$(\mathcal{L}_1) = 0.039$. Similarly, if we consider a less diverse subset including only Indo-European languages, e.g., $\mathcal{L}_2$ ={Italian, Polish, Russian} we have AbsGenDiv$(\mathcal{L}_2) = 1.531$ and GenDiv$(\mathcal{L}_2) = 0.017$. In this latter case, adding other Romance languages, e.g., Spanish, Catalan, and Portuguese, to $\mathcal{L}_2$ would increase GenDiv only to 0.022.

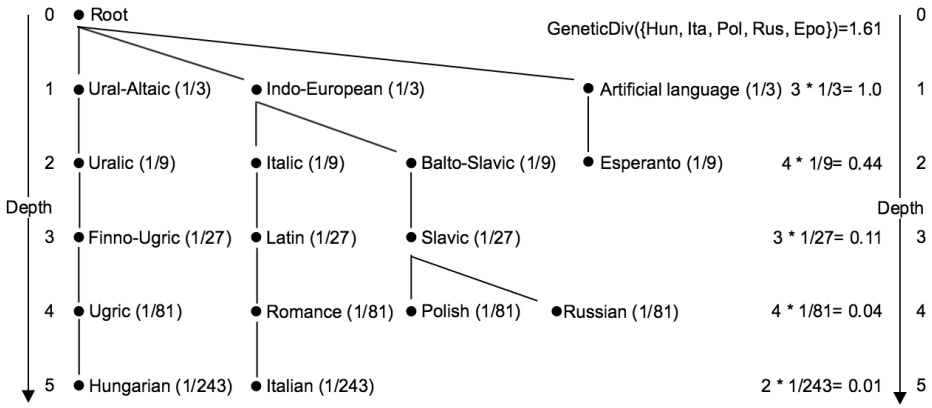

Figure 4.2: A example of absolute genetic diversity computation of related languages.

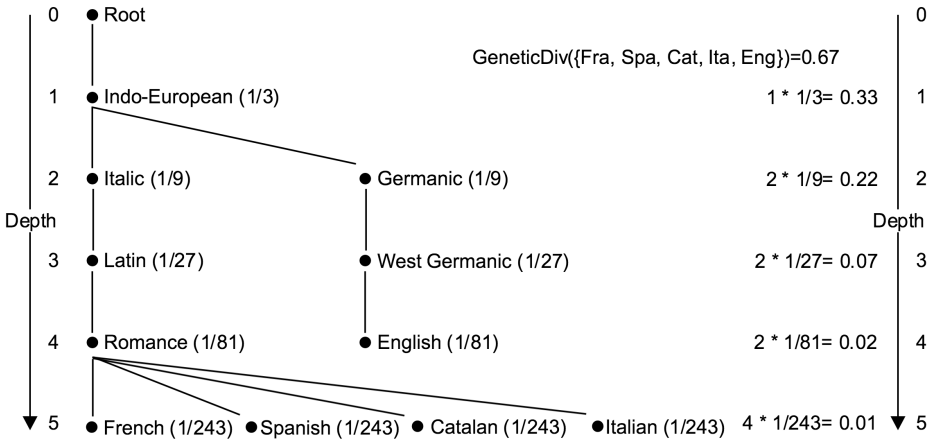

Figure 4.3: A example of absolute genetic diversity computation of unrelated languages.

4.1.2 Phylogenetic Tree of Language Families

In order to evaluate the quantitative measure of genetic diversity, we manually built the phylogenetic tree of language families by studying two language databases: WALS and Glottolog. The shortened version of the tree is displayed in Figure 4.4. As can be seen, each of second-level nodes in the tree represents an individual language family while a leaf node always accounts a individual language. Currently we built the tree storing 32 phyla and only 335 languages, designed to use UKC. The distribution of UKC languages in the tree is shown in Table 4.1.

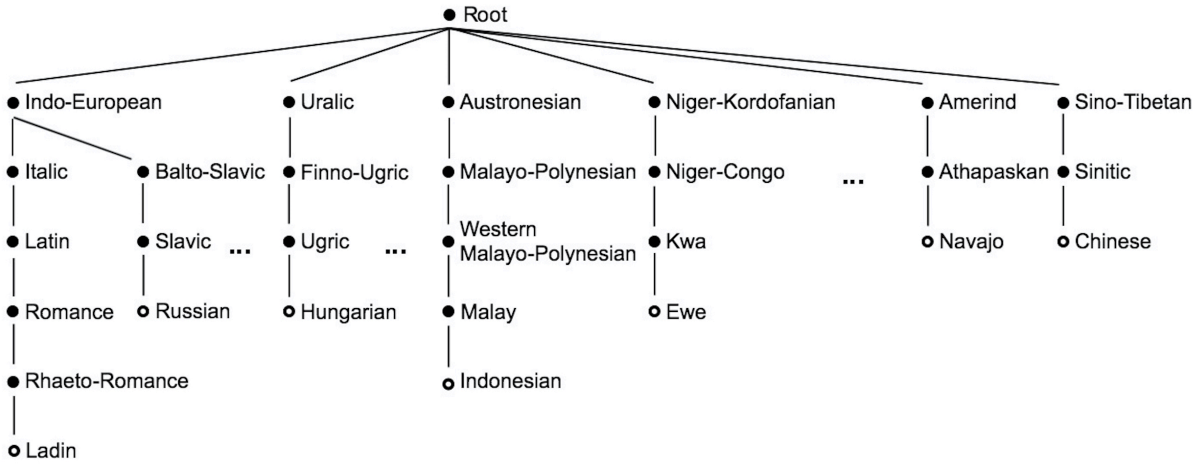

Figure 4.4: The phylogenetic tree of language families

4.2 Geographic Diversity

The definition of *geographic diversity* captures the intuition that languages with speakers living closely to one another tend to share more features and, in particular, a larger portion of their lexicon. This can be explained both diachronically (by the co-evolution of languages) and synchronically (these people will deal with the same types of objects and phenomena). As a first approximation, given that the UKC contains languages from everywhere in the world, we capture this

Table 4.1: Language distributions across phyla.

Phylum	**Depth**	**Languages**	**EU**	**AS**	**AM**	**AF**	**PA**	**Example**
Indo-European	7	115	86	26	1	1	1	English
Austronesian	6	36	1	23	2	0	10	Malay
Altaic	6	30	16	14	0	0	0	Mongolia
Uralic	6	22	22	00	0	0	0	Finnish
Niger-Kordofa.	5	21	0	0	0	21	0	Zulu
Amerind	4	18	0	0	18	0	0	Navajo
Sino-Tibetan	4	18	0	18	0	0	0	Mandarin
Afroasiatic	4	14	1	3	0	10	0	Hebrew
Caucasian	3	12	9	3	0	0	0	Chechen
Creole	3	9	0	0	5	1	3	Tok Pisin
small 22 families	4	40	4	11	17	4	4	Basque
Total	7	335	139	98	43	37	18	-

Depth represents a depth of its corresponding phylum.
Languages represents a number of languages existed in its corresponding phylum.
EU, **AS**, **AM**, **AF**, **PA** stand for continents, namely: Europe, Asia, Americas, Africa, and Pacific.
Note: each phylum of small families has no more than 5 languages.

intuition by defining our geographic diversity measure based on the number of different continents on which the languages in the reference data set are spoken. Then, the geographic diversity of $\mathcal{L}$ is taken to be 0 if $|\mathcal{L}| < 2$, and, otherwise, defined as:

$$\text{GeoDiv}(\mathcal{L}) = \frac{|\bigcup_{l \in \mathcal{L}} \text{continentOf}(l)|}{\#\text{Continents}} \tag{4.4}$$

where $\text{continentOf}(l)$ is the continent where l is spoken.

It is important to notice that the computation of geographic diversity through distance metrics alone is a gross over-simplification. Topology and the roughness of terrain, for instance, are important factors: mountain-dwelling people from geographically nearby valleys may in reality be completely isolated from each other. Historical periods of proximity are also ignored by synchronic only approaches, e.g., the temporary mixing of tribes having migrated together through

the Eurasian Steppe to then settle at great distances from each other. Still, at this stage, the values of diversity we compute are good enough to produce interesting results.

Chapter 5

Resource Quality

> "Quality is never an accident. It is always the result of intelligent effort. There must be the will to produce a superior thing." —-*John Ruskin*

The languages in the UKC are far from being *complete*, i.e., from containing all the words and synsets used in the everyday spoken or written interactions, and far from being *correct*, i.e., from containing only correct senses, namely, only correct associations from words and concepts to synsets. This situation is unavoidable. No matter how developed a language is, it will always miss a lot of words and it will always embody the misconceptions, bias and also mistakes of the people who have developed it. As mentioned in the introduction, in the area of historical linguistics, the solution so far has been that of using small high quality resources; see for instance the work in [McMahon and McMahon, 2005], in lexicostatistics [Swadesh, 1955, 1971], mass comparison [Greenberg, 1966], or the recent work on lexical semantics described in [Youn et al., 2016]. However this approach seems even more problematic as it does not give anyhow a full guarantee of unbiasedness, it tends to crystallize the field on a small set of case studies and, because of this, it makes it hard to study the diversity of languages *at large*, which seems to be a long tail phenomenon.

As from [Giunchiglia et al., 2017], our approach is to define a set of *quantitative measures* and use them to evaluate the quality of a language and of the bias it introduces. We believe that the quality of the resources is evaluated by two fundamental measures: 1. incompleteness and 2. incorrectness.

5.1 Incompleteness

Every description of lexical elements (e.g. word, sense, ...) in UKC has an incompleteness issue, so that we propose a number of measures for the incompletenesses of language, concept, and lexical ambiguity.

5.1.1 Quantifying Language Incompleteness

The proposed notion of *Language Incompleteness* LanInc, with its dual notion of *Language Coverage* LanCov, is the direct extension of the notion of incompleteness of logical languages and theories. The idea is to exploit the fact that the CC can be taken as (a computational representation) the domain of interpretation of a language, defined as a set of synsets, and to count how much of it is not lexicalized by that language.

$$\text{AbsLanCov}(l) = |\text{Concepts}(l)| \tag{5.1}$$

$$\text{LanCov}(l) = \frac{|\text{AbsLanCov}(l)|}{|\text{Concepts}(\text{UKC})| - |\text{Gaps}(l)|} \tag{5.2}$$

$$\text{LanInc}(l) = 1 - \text{LanCov}(l) \tag{5.3}$$

where $\text{Concepts}(l)$ is the set of concepts lexicalized by a language l, and $\text{Concepts}(\text{UKC})$ are the concepts in the UKC, and $\text{Gaps}(l)$ are the lexical gaps of l. AbsLanCov is the *Absolute Language Coverage*. Table 2 (column 10), reports the range of values for LangInc in the various phyla, while Table 3

Table 5.1: Language Groups.

Groups	Language Incompleteness	#Words	#Languages
a	LanInc$(l) \in [0.00; 0.52[$	$W \in [50,001; +\infty]$	6
b	LanInc$(l) \in [0.52; 0.82[$	$W \in [20,001; 50,000]$	15
c	LanInc$(l) \in [0.83; 0.99[$	$W \in [501; 20,000]$	64
d	LanInc$(l) \in [0.99; 1.00]$	$W \in [1; 500]$	250
UKC	LanInc$(l) \in [0.00; 1.00]$	$W \in [1; +\infty]$	335

Table 5.2: Ambiguity instances over the four language groups

Groups	Sample Languages	#AmbIns	AvgAmbCov
a	English, Finnish, ...	714,437	10.2
b	Dutch, Spanish, ...	1,969,436	12.8
c	Zulu, Tswana, ...	117,213	18.4
d	Ewe, Abakhaz, ...	1,725	35.5
UKC	–	2,802,811	12.4

provides its values for ten selected languages. It is interesting to notice how LangInc(English) $= 0.0$. This is indirect evidence of the English bias present in the current linguistic resources. It is a consequence of the fact that most Wordnets have been derived by PWN and that, so far, the UKC contains only concepts lexicalized in the PWN. The second observation is that all the languages not spoken by WEIRD societies are highly under-developed, for instance we have LangInc(Navajo) $= 0.98$.

Table 5.1 (left) shows the level of incompleteness of the various languages of the UKC. This table organizes languages into four groups where groups (a), (b) are highly developed while groups (c), (d) are highly under-developed.

5.1.2 Quantifying Concept Incompleteness

The notion of *concept incompleteness* can be thought of as the dual of language incompleteness. If the latter measures how much of the UKC a language does not cover, the former measures how much a single concept is covered across a

selected set of languages. Let, for any concept c, the *Languages of* c be the set of languages where c is lexicalized, defined as:

$$\text{Languages}(c) = \bigcup_{l \in \mathcal{L}} \{l | \sigma(c, l) > 0\} \tag{5.4}$$

where $\sigma(c, l)$ returns either 1 or 0, depending on whether c is lexicalized in l. Then we define *concept coverage* and of *concept incompleteness* as follows:

$$\text{AbsConCov}(c) = |\text{Languages}(c)| \tag{5.5}$$

$$\text{ConCov}(c) = \frac{\text{AbsConCov}(c)}{|\text{Languages}(\text{UKC})|} \tag{5.6}$$

$$\text{ConInc}(c) = 1 - \text{ConCov}(c) \tag{5.7}$$

In words: the *absolute coverage* of a concept is the cardinality of the set of languages where it occurs, its coverage is the absolute coverage normalized over the number of languages of the UKC (defined as Languages(UKC) with a slight abuse of notation), its incompleteness is the complement to 1 of its coverage.

5.1.3 Quantifying Ambiguity Incompleteness

Figure 5.1 shows the distribution of concepts for each value of AbsConCov(c) with a concept c standing for the set of the concepts corresponding to the four parts of speech (i.e., adjective, adverb, noun, and verb). As it can be seen from the mean line, on average, concepts are lexicalised across about 12.93 languages.

The notion of *ambiguity incompleteness* integrates the notion of language incompleteness. As it is well known, the key difference between logical languages and natural languages is that the latter, differently from the former, allow

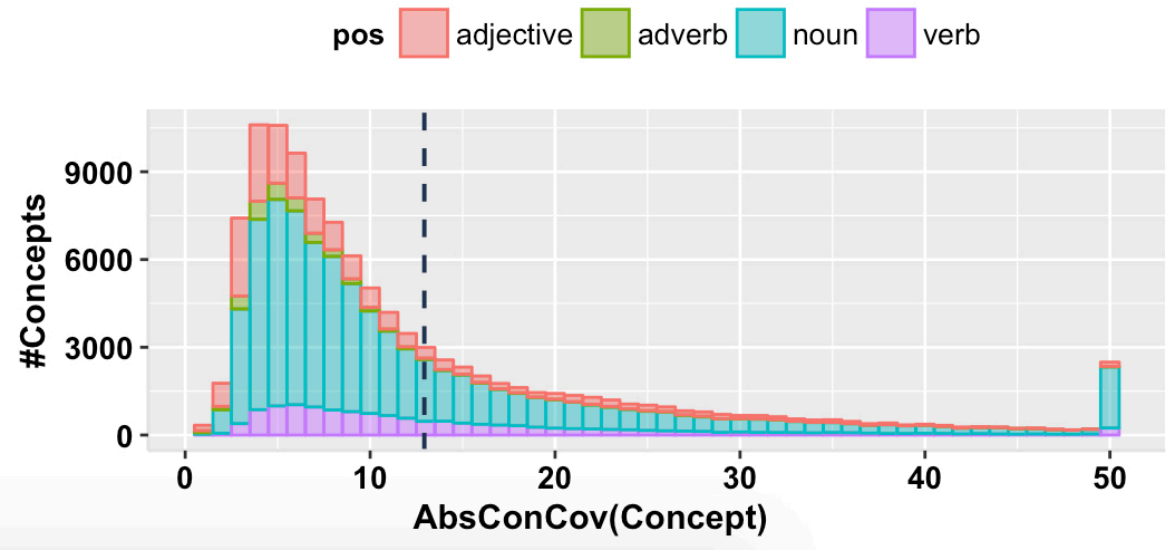

Figure 5.1: Concept distributions per AbsConCov value.

words to denote more than one concept. This phenomenon gives rise to the phenomenon of lexical ambiguity, e.g., polysemy or homonymy. Let the 4-tuple a =<w, c_1, c_2, l >be an *ambiguity instance*, where c_1 and c_2 are two concepts expressed by the same word w in the language l. We define *ambiguity coverage* and *ambiguity incompleteness* as follows:

$$\text{AbsAmbCov}(a) = |\text{Languages}(c_1) \cap \text{Languages}(c_2)| \tag{5.8}$$

$$\text{AmbCov}(a) = \frac{\text{AbsAmbCov}(a)}{|\text{Languages}(\text{UKC})|} \tag{5.9}$$

$$\text{AmbInc}(a) = 1 - \text{AbsAmbCov}(a) \tag{5.10}$$

In words: the *absolute ambiguity coverage* of a word together with its two concepts is the number of languages where these concepts occur, its coverage is the absolute coverage normalized over the number of languages of the UKC, its incompleteness is the complement to 1 of its coverage.

Let AmbInstances($\mathcal{L}$) be the *set of ambiguity instances* in the set of languages $\mathcal{L}$ and AvgAbsAmbCov($\mathcal{L}$) the *average absolute ambiguity coverage*, which we compute as follows:

$$\frac{\sum_{a \in \text{AmbInstances}(\mathcal{L})} \text{AbsAmbCov}(a)}{|\text{AmbInstances}(\mathcal{L})|} \tag{5.11}$$

Table 5.2 (right) reports the number of ambiguity instances and their average number for the four language groups plus the UKC. Notice how the average absolute ambiguity coverage is much higher for the under-developed language groups (c), (d). In other words language coverage increases when the average ambiguity coverage decreases, and vice versa: the more developed a resource is the less ambiguity instances we have. This fact, counter-intuitive at a first sight, is a consequence of the fact that, in practice, the first words added to a language are the ones which are most commonly used and therefore, the most ambiguous.

Table 5.3: The most polysemous ten words in UKC

	ISO	**Languages**	**Lemma**	**Senses**
1	slv	Slovene	biti	701
2	slv	Slovene	imeti	164
3	msa	Malay	membawa	130
4	ind	Indonesian	membawa	130
5	msa	Malay	membentuk	107
6	ind	Indonesian	membentuk	107
7	fra	French	donner	102
8	fra	French	faire	101
9	ind	Indonesian	membuat	100
10	msa	Malay	membuat	100

5.2 Incorrectness

The correctness of a language can be measured by several factors, e.g., translation mistakes, wrong senses, and much more. In particular, the state-of-the-art multilingual lexical resources like BabelNet and Open Multilingual Wordnet integrates many wordnets from several different sources while many of them were semi-automatically or fully-automatically built, and the rest of them in general employed a method of translating PWN to their target languages. As a result, all their issues and mistakes were migrated to the bigger resource like Babelnet. For instance, PWN is organized under the psycholinguistic principles based on the research of English language and psycholinguistic studies while wordnets in other languages tends to forgot those fundamental principles. And other example is a result of automatic integration method of dictionaries.

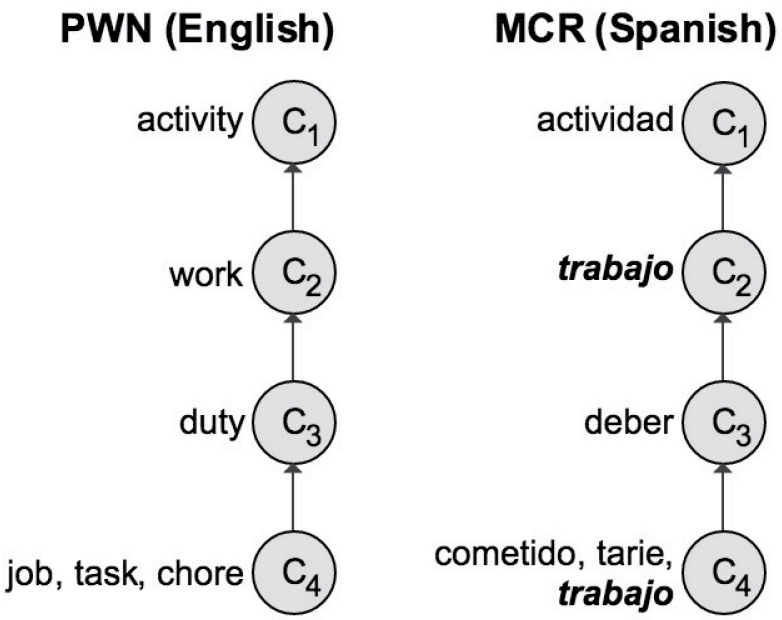

Figure 5.2: A psycholinguistic mistake in Spanish.

Table 5.4: Ten sample languages from ten phyla in Table 4.1.

Language	ISO	#PsyMis	AvgDis	LanInc	LanQua
English	eng	14	3.42	0.00	1.00
Malay	msa	4,304	1.46	0.71	0.16
Mongolia	mon	6	1.16	0.99	0.50
Finnish	fin	7,471	1.22	0.01	0.27
Ewe	ewe	0	0	0.99	0.59
Navajo	nav	54	1.44	0.98	0.37
Mandarin	zho	2,596	1.17	0.09	0.38
Hebrew	heb	49	1.23	0.33	0.43
Chechen	che	0	0	0.99	0.61
Tok Pisin	tpi	22	1.68	0.99	0.28

5.2.1 Quantifying Language Incorrectness

In the following, we analyze the problem of the *psycholinguistic mistakes* which we define as failures of adhering to the principle which, as from [Miller, 1990], states that *"... superordinate nouns can serve as anaphors referring back to their hyponyms. For example, in such constructions as 'He owned a rifle, but the gun had not been fired', it is immediately understood that the gun is an anaphoric noun with a rifle as its antecedent."* Figure 5.2 provides an example of psycholinguistic mistake in the Spanish WordNet. We have the following definitions:

$$\text{AbsLanQua}(l) = -\log_{10}(\frac{|\text{PsyMis}(l)| + 1}{|\text{Concepts}(l)|}) \tag{5.12}$$

$$\text{LanQua}(l) = \frac{\text{AbsLanQua}(l)}{\text{AbsLanQua}(\text{English})} \tag{5.13}$$

$$\text{AvgDis}(l) = \frac{\sum_{x \in \text{PsyMis}(l)} \text{dis}(x)}{|\text{PsyMis}(l)|} \tag{5.14}$$

where $\text{PsyMis}(l)$ is the set of psycholinguistic mistakes in l, $\text{AbsLanQua}(l)$ and

LanQua(l) are the *Absolute Language quality* and the *Language quality* of *l*, respectively. The number of mistakes varies a lot, going from the fourteen mistakes of the PWN English to the thousands of mistakes of other languages such as Chinese and Finnish. The Log-based definition of AbsLanQua is meant to alleviate this problem (see Tables 2 and 3). English is taken to be the reference to which we normalize the quality of the other languages. $dis(\mathrm{x})$ is the number of intermediate nodes between two concepts generating the psycholinguistic mistake x, for instance, in figure 5.2, $dis(\text{trabajo}) = 2$. The *Average Distance* AvgDis measures the average distance for a language. As from Table 3, this distance is around 1 for most languages with the exception of English where it is 3.42, which provides even more evidence of the large gap in quality between the PWN English and any other language.

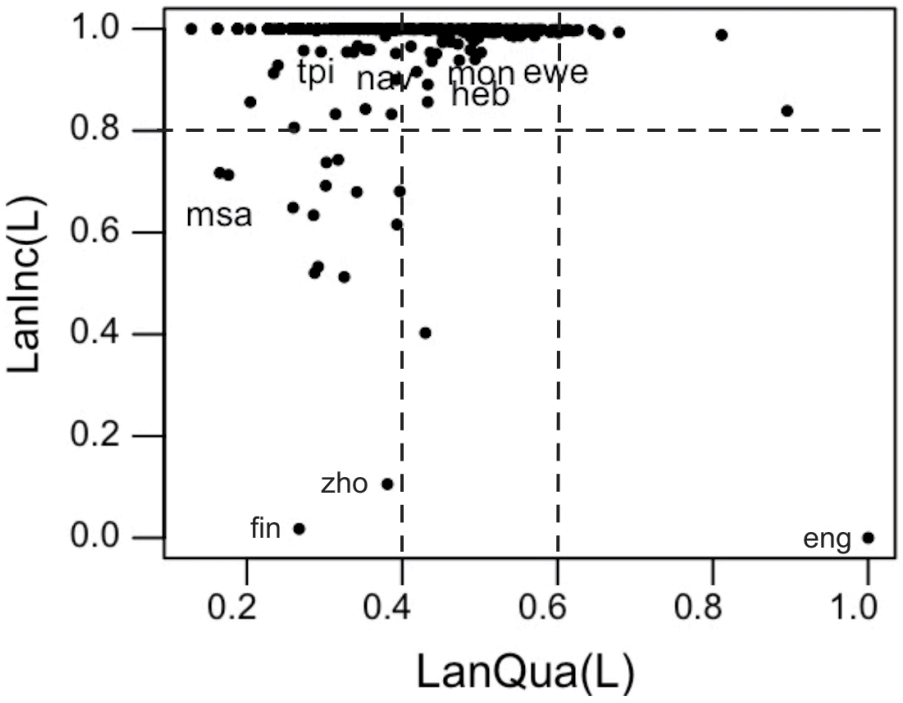

Figure 5.3: Language Incompleteness vs Language Quality.

5.2.2 Incorrectness distribution of UKC languages

Figures 5.3 and 5.4 compare the incompleteness and quality values of the languages in the UKC, where the ten languages in Table 3 are explicitly marked with their ISO names, as from Table 3.

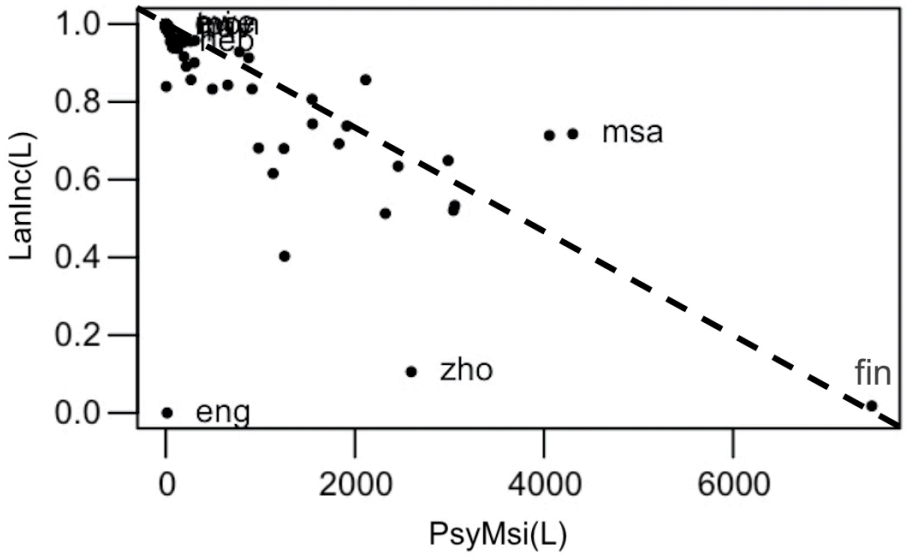

Figure 5.4: Language Incompleteness vs Psycholinguistic Mistakes.

Figure 5.3 shows that most languages have a low quality, below 0.4, and that the most developed languages (the ones with LanInc below 0.7), with the exception of English, have even lower values. In other words, there are only few languages which are highly developed but they have much poorer quality (<0.42) than English (=0.99). It basically means that none of those lexical resources employed the psycholinguistic principles, given by Miller.

Figure 5.4 compares incompleteness and the absolute number of mistakes. Here, the majority of languages is below the dashed line making even more explicit how the number of mistakes grows with the size of the resource.

5.3 Summary

In this chapter, we illustrated the sets of quantitative measures of resource incompleteness and quality. By using those measures, we showed that the current trending nature of quality bias towards the wordnets except for PWN.

Chapter 6

Polysemy vs Homonymy

> "If a word exhibits polysemy in one language, one may be inclined, or forced, to dismiss its various meanings as coincidental; if a corresponding word in another language exhibits the same, or closely parallel polysemy, it becomes an extremely interesting coincidence; if it displays the same polysemy in four, five, or seven genetically unrelated languages, by statistical law it ceases to be a coincidence at all." —-*John Haiman, 1974*

6.1 Lexical Diversity in Semantic Relatedness

The issue of *Lexical Semantic Relatedness* has been extensively studied, see, e.g., [Budanitsky and Hirst, 2006]. However, all the work so far has mainly, if not exclusively, concentrated on its study within a single language while we focus on how semantic relatedness propagates across languages. To get an insight into the problem, consider the three examples in Tables 6.1, 6.2, 6.3. These tables provide examples of the types of semantic relatedness we consider. Notice that we distinguish between two types of morphological relatedness: *compounding*,[1], namely the combination of free morphemes (as in $key + board \rightarrow keyboard$),

[1]We use the term *compounding* to cover also idioms and collocations where component words are separated by spaces: *hot dog*, *tax cut*. This is justified by the fact that the presence or absence of spaces is more a matter of

Table 6.1: An example of polysemy in English.

#	Language	Concept 1	Concept 2	Types
1	English	bar	bar	polyseme
2	Italian	barra	bar	derivational
3	Mongolian	тээк	баар	different
4	Chinese	酒吧	酒馆	derivational
…	…	…	…	…
23	Finnish	baaritiski	baari	compound
Types	**polyseme**	**compound**	**derivational**	**different**
Languages	11	1	5	6

Concept 1: a counter where you can obtain food or drink.

Concept 2: an establishment where alcoholic drinks are served over a counter.

Table 6.2: An example of homonymy in English.

#	Language	Concept 1	Concept 2	Types
1	English	melody, air	air	homonym
2	Italian	melodia, aria	aria	homonym
3	Mongolian	аялгуу	агаар	different
4	Chinese	旋律	空气	different
…	…	…	…	…
38	Turkish	melodi	hava	different
Types	**homonym**	**compound**	**derivational**	**different**
Languages	6	0	0	32

Concept 1: a succession of notes forming a distinctive sequence.

Concept 2: a mixture of gases (especially oxygen) required for breathing.

and *derivation* namely the combination of a word with one or more derivational affixes (bound morphemes) (as in *play* + *-er* $\rightarrow$ *player*).

The key observation is that diverse languages represent the same semantic relatedness in diverse ways. Thus, for instance, in Table 6.1, a polyseme in English corresponds to an occurrence of derivational morphology in Italian and Chinese, to an occurrence of compound morphology in Finnish and to two distinct words

language-specific orthographical convention than a semantic differentiator (e.g., English prefers multiword expressons, German tends to use compounding, whereas some languages such as Chinese do not use spaces to separate words at all).

Table 6.3: An example of compound morphology in English.

#	Language	Concept 1	Concept 2	Types
1	English	tennis	tennis player	compound
2	Italian	tennist	tennista	derivational
3	Mongolian	теннис	теннисчин	derivational
4	Chinese	网球	网球选手	compound
…	…	…	…	…
25	Korean	테니스	테니스선수	compound
Types	**polysemy**	**compound**	**derivational**	**different**
Languages	0	11	14	0

Concept 1: a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court.

Concept 2: an athlete who plays tennis.

in Mongolian.

6.2 Method

Our goal is to establish whether any two concepts denoted by a single word are polysemes or homonyms. The algorithm we propose is based on the following intuitions:

- if two concepts are semantically related in diverse languages, then they are polysemes. In this case the diversity of the two languages is evidence of the fact that semantic relatedness derives from a property of the world, which is what all languages denote.

- if two concepts are *not* semantically related in diverse languages, then they are homonyms. The key idea is that the occurrence of a homonym in a single language, or in similar languages is a coincidence, a consequence of some local, e.g., contextual or cultural, phenomena.

- Similar languages provide little support for the discovery of polysemes and

homonyms. At the same time, the existence of polysemes and homonyms can be propagated across similar languages.

Algorithm 1: Lexical Ambiguity Classification

Input : $x =<l, w, c_1, c_2>$, an ambiguity instance
Input : $\mathcal{R}$, a multilingual lexical resource
Output : $label$, an ambiguity class for the instance a.

1 $\mathcal{L}_P \leftarrow \emptyset$;
2 $\mathcal{L} \leftarrow \text{Languages}_{\mathcal{R}}(c_1) \cap \text{Languages}_{\mathcal{R}}(c_2)$;
3 **for** *each language* $l \in \mathcal{L}$ **do**
4 **for** *each word* $w_1 \in \textit{Words}_{\mathcal{R}}(c_1, l)$ **do**
5 **for** *each word* $w_2 \in \textit{Words}_{\mathcal{R}}(c_2, l)$ **do**
6 **if** $w_1 = w_2$ ***or*** $\textit{morphSim}(w_1, w_2)$ **then**
7 $\mathcal{L}_P \leftarrow \mathcal{L}_P \cup \{l\}$;
8 $\mathcal{L}_H \leftarrow \mathcal{L} - \mathcal{L}_P$;
9 **if** $\textit{ComDiv}(\mathcal{L}_P) > T_D$ **then**
10 $label \leftarrow$ 'polyseme' ;
11 **else if** $\textit{ComDiv}(\mathcal{L}_H) > T_D$ ***and*** $\textit{ComDiv}(\mathcal{L}_P) < T_S$ **then**
12 $label \leftarrow$ 'homonym' ;
13 **else**
14 $label \leftarrow$ 'unclassified' ;
15 **return** $label$;

But, how do we automatically recognize that two concepts are semantically related? The idea is simple: if we have a big enough number of diverse languages where the two words denoting the two concepts are syntactically similar, then the two concepts are semantically related. A consistent use of the similar words is evidence of semantic relatedness, as it also the case in the examples in Tables 6.1 and 6.3. The resulting algorithm (see algorithm 1) takes in input an ambiguity instance x and a multilingual resource and it returns one of three classifications for x: *polyseme*, *homonym* or *unclassified*. This algorithm is structured as follows:

Step 1. (Lines 1-2). It initializes the set $\mathcal{L}_P$ of the languages supporting the

occurrence of a polyseme (Line 1) and it collects in $\mathcal{L}$ all the languages where c_1 and c_2 are lexicalized (Line 2);

Step 2. (Lines 3-7). It tries to recognize x as a candidate polyseme. This attempt succeeds if one of two conditions hold: (i) the two words are the same, i.e., we have discovered another case of polisemy in a new language or (ii) the two words are *morphologically related*, as computed by the function morphSim. If it succeeds it adds l to $\mathcal{L}_P$.

$$\text{morphSim}(w_1, w_2) = \frac{\text{len}(\text{LCA}(w_1, w_2))}{\max(\text{len}(w_1), \text{len}(w_2)))} \tag{6.1}$$

Our current implementation of morphSim, is a (quite primitive) string similarity metric. For w_1 and w_2 to be related, morphSim(w_1, w_2) must return a value higher than a threshold T_M. The function len() returns the length of its input while the function LCA() returns the *longest common affix* (prefix or suffix) of the two input words: for example, '*compet*' is the LCA for the words '*compete*' and '*competition*'.

Step 3. (Line 8) It creates the set $\mathcal{L}_H$ of the languages supporting the occurrence of a homonym. Notice how $\mathcal{L}_H$ contains the languages where w_1 and w_2 are different words.

Step 4. (Lines 9-14) x is classified. Notice that, for x, to be classified as a polyseme, the combined diversity of $\mathcal{L}_P$ must be higher than T_D (where "D" stands for Diversity) while, to be classified as a homonym, the combined diversity of $\mathcal{L}_H$ must be higher than T_D and lower than T_S (where "S" stands for Similarity). We call T_D and T_S the *Diversity Threshold* and the *Similarity Threshold*, respectively. The intuition is that an ambiguity instance is a polyseme if it occurs in a "diverse enough" language set while it is a homonym if it occurs in a language set where the languages supporting homonymy are "diverse enough"

and the languages supporting polisemy are "similar enough". One such example are the two homonyms, one in English and one in Italian, in Table 6.2.

6.3 Results

We organize this section in three parts. First we describe how we have learned the hyperparameters. Then we describe the results of the experiment. Finally we analyze the impact of incompleteness on the experiment itself.

Table 6.4: Parameter configuration and comparisons.

	Homonym			Polyseme		
Methods	**Recall**	**Precision**	**F1**	**Recall**	**Precision**	**F1**
Baseline	59.58	58.00	58.77	17.64	100	29.98
Rijkhoff	12.71	95.65	22.44	11.56	95.23	20.61
AbsGenDiv	15.6	96.42	26.86	26.01	95.74	40.9

Baseline: no parameters.
Rijkhoff: $\beta = 1.4, T_D = 47.2, T_S = 13.2, T_M = 0.5$.
AbsGenDiv: $\beta = 1.0, T_D = 2.52, T_S = 0.68, \lambda = 2.7, T_M = 0.5$.

6.3.1 Algorithm Configuration

The hyperparameters to be identified are: the weight β of geographic diversity with respect to genetic diversity, the parameter λ for the computation of genetic diversity, the diversity threshold T_D and the similarity threshold T_S.

We have computed these values in two steps. First, we have selected a grid of value configurations. The grid has been built by taking, for each parameter, an increment of 0.1 within the following ranges: $\lambda = [1.2; 4.0]$ (higher values favour more phyla in the language set), $T_D = [1.0, 10.0]$ (the higher the value the more diversity is required for polysemy and homonymy detection), $T_S = [0.3, 1.7]$ (the lower the value the more similarity is allowed for homonymy),

$\beta = [0.0; 1.5]$ (the lower the less relative significance of geographic diversity), $T_M = [0.5, 0.8]$. The number of configurations which have been analyzed is: 28 (variations on λ) $\times$ 90 (variations on T_D) $\times$ 15 (variations on T_S) $\times$ 16 (variations on β) $\times$ 4 (variations on T_M) = 2,419,200 configurations.

Then we have run algorithm 1 with three different methods for computing genetic diversity namely, AbsGenDiv (and not GenDiv: while being conceptually the same, it produced values for β less close to 0), the measure defined in [Rijkhoff et al., 1993] and Baseline, a simple greedy algorithm where an ambiguity instance is classified as a polyseme if $\mathcal{L}_+$ contains at least 3 phyla and as a homonym if $\mathcal{L}_+$ contains only 1 phylum. In all three cases we have learned the parameters (λ, β, T_D, T_S, T_M) using a training set of 173 polysemes and 146 homonyms from three phyla. Since our ultimate goal is to generate high-quality knowledge, we have favoured precision over recall, setting our minimum precision threshold to 95% and maximising recall with respect to this constraint. The best settings as well as the corresponding precision-recall figures, as computed on the training set, are reported in Table 6.4. As it can be seen, AbsGenDiv is uniformly better than Rijkhoff's and only loses to Baseline on the recall of homonym classification, which is not relevant, given our focus on precision.

6.3.2 Polysemy vs. Homonymy

The UKC contains 2,802,811 ambiguity instances across its pool of 335 languages, These instances were automatically generated and then given in input to the algorithm which, in turn, generated 908,110 candidate polysemes and 594,115 candidate homonyms across all languages.

A sample of 640 cases, half being candidate homonyms and half being candidate polysemes, were randomly selected, which were equally divided across seven languages belonging to six different phyla (English, Hindi, Hungarian, Korean, Kazakh, Chinese, Arabic). Seven native speakers were selected as evaluators. All the evaluators, though not being linguists by training, had previously

Table 6.5: Evaluated precision on polysemes and homonyms

				Precision%	
Languages	**Polysemes**	**Homonyms**	**Samples**	**Homo.%**	**Poly.%**
English	34,625	10,551	100	48	99
Kazakh	34	6	40	66	97
Hungarian	1,284	246	100	44	100
Hindi	342	57	100	92	98
Chinese	16,450	5,481	100	61	100
Korean	542	260	100	46	98
Arabic	7,973	3451	100	26	100
Average				52.1	98.5

had some exposure to *WordNet*. They were provided with the glosses of the concepts involved, they were asked the follwing question: "Do you think meanings c_1 and c_2 of word w are related?", and they had to provide a yes/no answer.

Table 6.5 provides statistics and accuracy values for each of the languages evaluated. The average accuracy for finding polysemes is 98.3%, even higher than with the training set. Our explanation is that the evaluation dataset is more diverse than the training dataset, as it contains languages from six phyla instead of three. The accuracy of homonym detection is much lower (52.2%), but still significantly higher than what one would obtain by random guessing. At the moment it is unclear whether this lower accuracy is because there are many cases of occurrences of what we call *isolated polysemes*, namely polysemes occurring in a single language (or a set of similar languages) or, more simply, a consequence of the incompleteness of the UKC. It is a fact that accuracy grows substantially if one increases the number of ambiguity instances considered (see next section). This is a topic for future investigation.

6.3.3 The Impact of Resource Incompleteness

We have organized this study following the various steps of the algorithm. Table 6.6 shows how resource incompleteness impacts the computation of ambiguity

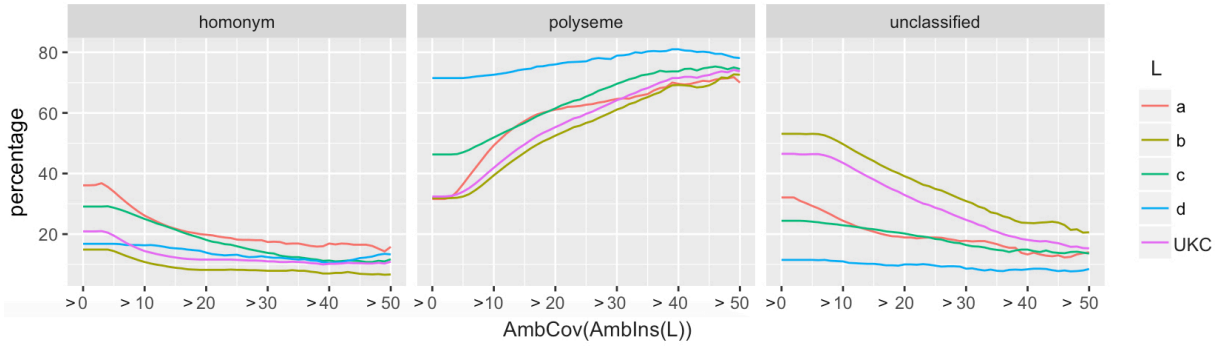

Figure 6.1: Classification results vs. required minimal number of ambiguity instances.

instances. It does it in three parts (the three main rows): first by incrementally increasing the languages being analyzed (by adding language groups), then by analyzing the 4 language groups one by one, and finally by analyzing some reference languages. The Tasks column reports the languages being analyzed (thus, for instance (a+b) means all the languages in groups (a) and (b). The Resource column reports the resource over which the analysis is performed. Thus, the first group corresponds to the case where all the languages in the resource are considered; the second group corresponds to the case where the languages in a group are studied in the UKC (namely (a+b+c+d)) while the last group corresponds to the study of single languages in the UKC. The third column provides the classification results.

The overall results show various facts: (i) from the first column, the number of ambiguity instances grows with the size of the languages considered (namely with the total number of words in a language set), as it should be expected; (ii) from the second column, the average number of ambiguity instances increases with the decrease of language coverage also for single languages, thus confirming what discussed in Section 4 (and reported in this table in the second row of this column); (iii) the number of unclassified cases is quite high and decreases with the decrease of the overall language coverage (see second row; remember that group (b) contains many more languages that group (a), see Table 5.1), which seems coherent with the previous observation.

Table 6.6: Language coverage and classification results.

Tasks		Resource		Classification Results		
Groups	**#AmbIns**	**Groups**	**AvgAmbCov**	**Poly.%**	**Homo.%**	**Uncl.%**
a	714,437	a	4.19	13.0	43.9	43.1
a+b	2,683,873	a+b	10.89	30.8	21.1	48.1
a+b+c	2,801,086	a+b+c	12.40	32.4	21.2	46.4
a+b+c+d*	2,802,811	a+b+c+d	12.43	32.4	21.2	46.4
a	714,437	a+b+c+d	10.28	31.6	36.1	32.1
b	1,969,436	a+b+c+d	12.83	31.9	14.9	53.1
c	117,213	a+b+c+d	18.47	46.3	29.1	24.4
d	1,725	a+b+c+d	35.51	71.5	16.8	11.5
English (a)	197,502	a+b+c+d	9.67	32.2	22.9	44.7
Slovene (b)	156,317	a+b+c+d	12.18	35.5	27.0	37.4
Hungary(c)	1,907	a+b+c+d	21.67	65.7	14.9	19.2
Haitian (d)	39	a+b+c+d	29.69	87.1	5.1	7.6

* a+b+c+d = UKC.

Table 6.6 links thee average number of ambiguity instances with the classification results. Figure 6.1 refines this results by showing how, limited to the language groups (a), (b), (c), (d), and the UKC (as reported in the middle of Table 6.6), the minimal number of ambiguity instances ($> 0, > 10, > 20, ...$) which are required for accepting an ambiguity instance as such, impacts the classification results. It shows how, for all the language groups, with the growth of the minimal number of required ambiguity instances, the proportion of homonyms

Table 6.7: UKC classification results from Figure

UKC		Classification Results		
AmbCov	**#AmbIns**	**Polyseme%**	**Homonymy%**	**Unclassified%**
>0	2,802,811	32.4	21.2	46.4
>10	1,805,144	41.9	14.4	43.5
>20	325,322	55.3	11.6	32.9
>30	44,408	64.2	11.0	24.7
>40	9,556	71.5	10.2	18.1
>50	3,198	73.7	10.9	15.3

Table 6.8: Classification accuracy vs. ambiguity coverage.

				Accuracy%	
AmbCov	**#Polysemes**	**#Homonyms**	**Total**	**Hom.%**	**Pol.%**
>0	334	306	640	52.2	98.3
>10	267	297	564	52.9	98.5
>20	173	143	316	60.1	98.8
>30	103	33	136	69.7	99.0
>40	56	10	66	70.0	98.2
>50	30	7	37	71.4	100.0

tends to converge to a low percentage (below the 20%), while the proportion of polysemes tends to converge to a very high percentage (above the 70%), and the proportion of unclassified instances decreases substantially (below the 20%). This is coherent with our expectation of a very low percentage of honymyms, most likely below the 10%.

Table 6.7 provides the numeric quantification of the UKC results graphically represented in Figure 6.1, together with the extra information of the number of instances computed. It can be noticed how increasing the minimal required number of ambiguity instances consistently increases the percentage of polysemes (up to the 73.7%), decreases the percentage of homonyms (down to the 10.9%) as well as the percentage of unclassified instances (down to around the 15.3%)

Table 6.8 refines the results in Table 6.7 by showing how the accuracy with polysemes and homonyms grows with the growth of AmbCov, namely with the growth of the number of languages where the two concepts occurring in an ambiguity instance are lexicalized. It can be seen the accuracy of polysemy is very robust while that of homonymy is highly sensitive to the number of languages, converging to high levels of accuracy.

6.4 Summary

In this chapter we have presented a general approach which allows to use large

scale resources, in our case, the UKC, to solve relevant problems in linguistics and use the results to improve the UKC itself. The proposed approach has been applied to the discovery of homonyms, as distinct from polysemes, in the UKC. Our current work is concentrated on developing other case studies and on using them to validate and refine the proposed methodology.

Chapter 7

Discovery of Lexical Relations

"Language is a city to the building of which every human being brought a stone." —-*Ralph Waldo Emerson*

The Princeton WordNet is organized by the relations between word meanings or word forms. Those relations are divided into two types, namely: lexical and semantic relations. The semantic relations are language independent, so that both in principle and practice the semantic relations are easily incorporated into the lexical resources, aligned with PWN. In contrast to the fact just mentioned, the lexical relations are not imported to new languages due to language diversity. Consequently, the lexical relations are being either manually or semi-automatically developed in other WordNets such as MultiWordNet [Pianta et al., 2002a].

7.1 Backgrounds for Lexical Relations

The UKC has the five types of lexical relations as follows.

1. "Pertain" is a lexical relation between the relational adjective and a noun that the adjective is pertaining to (e.g. chemical is pertaining to chemistry).

Sometimes, it also is about a lexical relation between the relational participle adjective and a verb that is derived from the adjective (e.g. chemically is derived from chemical).

2. "Antonymy" is a lexical relation between word forms, not a semantic relation between word meanings. For example the meanings {*rise, ascend*} and {*fall, descend*} are conceptual opposite, but they are not antonym; [*rise, fall*] are antonyms and so are [*ascend, descend*].

3. "Derived" is a lexical relation in the terms of that word forms in different syntactic categories that have the same root form and are semantically related (e.g. bassoon and bassoonist).

4. "Homonymy" is a lexical symmetric relation used to explicitly mark two senses of same word form and part-of-speech having unrelated meanings. A classic example is the word "bank" as institution and bank as sloping land.

5. "Part-of" is a lexical symmetric relation used to explicitly mark two senses of same word form and part-of-speech having closely related meanings. A classic example is the word "university" as institution and university as building.

The majority of WordNet-like lexical resources in other languages have very little information about lexical relations while NLP applications such as Information Retrieval and Machine Translation rely critically on it.

Recently, a number of semi-automatic methods [Pala and Hlaváčková, 2007] [Koeva, 2008] [Fellbaum et al., 2007] have been proposed to build the "derived" lexical relations by studying the morphological derivation which is the process of forming a new word on the basis of an existing word by adding the derivational morphemes, e.g. happiness and unhappy from the root word (base words) happy.

In general, the authors made the list of the derivative affixes by asking from the linguistics or using the existing resources. Then, those derivational morphomes were used to generate the candidate pairs of words from the existing WordNets. Finally, those candidate pairs were validated manually. In this way, however, building the lexical relations manually requires a lot of years due to the large number of languages. Therefore, in the following subsection, we propose the automatic method to discover the lexical relations.

7.2 Method

One interesting solution to this problem is the cross-lingual analysis to automatically discover the lexical relations, that are based on derivational morphology. In other words, the lexical relations in PWN give a signal that same lexical relations between the exact same concepts could be existed in other languages. The example between English and Spanish is shown in Figure 7.1.

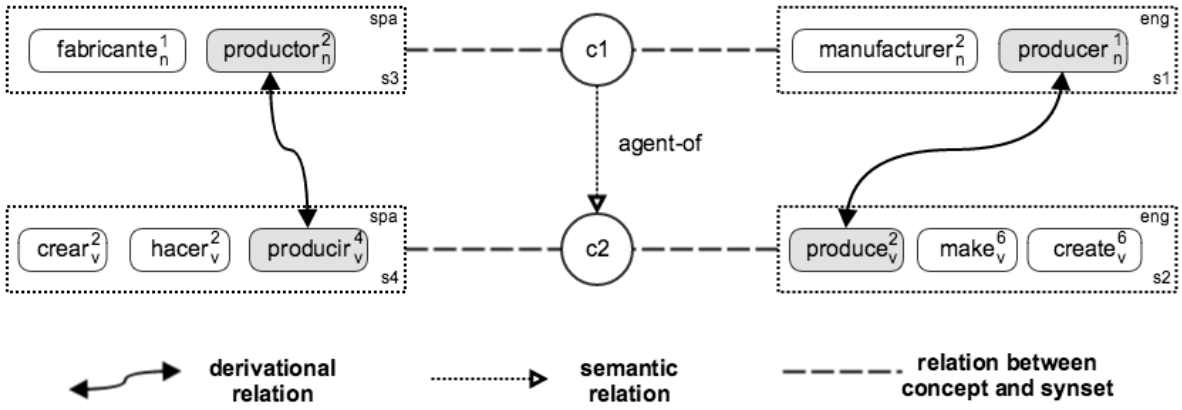

Figure 7.1: The cross-lingual example of the derivational relation in Spanish and English

Our proposed method is described in Algorithm 1 that takes as input as R, the set of lexical relations in PWN and L, the set of languages, and returns as output as S, the set of discovered lexical relations in the given languages. In the algorithm, for the steps 2-6 each of lexical relations in PWN is first being

Table 7.1: Examples of lexical relations with LCA and DM values

Languages	$word_a$	$word_b$	**relation**	**LCA**	**DM**
Italian	zia (aunt)	zio (uncle)	antonym	zi	0.66
Spanish	humear (smoking)	humo (smoke)	derived	hum	0.5
Hindi	पर्यटन (tourism)	पर्यटक (tourist)	derived	पर्यट	0.8
Chinese	幼虫 (larva)	幼虫的 (larval)	pertain	幼虫	0.66
Hungarian	értékes(valuable)	értéktelen(worthless)	antonym	érték	0.5

iterated, then each of the given languages is being iterated to investigate whether the lexical relation *r* in the language *l* exists between synsets syn_1 and syn_2. At the step 7, both synsets syn_1 and syn_2 are being checked as null value. If either one is null, the steps 8-13 are skipped. Otherwise, each possible combination of sense pairs of the synsets are being validated whether it qualifies the distance metric threshold (Eq. 1). If *sim* for the candidate sense pair is greater than the threshold β, the lexical relation is created between the sense pair and added to *S* the set of lexical relations at the step 13. Finally, *S*, the set of all the discovered lexical relations, is returned at the step 14.

Distance Metric. It is a string metric for measuring the similarity between two strings based on the longest common affix (LCA). The value range is between 0 and 1. If it is closer to 1, the two strings are very common in either prefix or suffix.

$$DM = \frac{Len(LCA(w_a, w_b))}{max(Len(w_a), Len(w_b)))} \tag{7.1}$$

where w_a and w_b are the given words, and $LCA(w_a,w_b)$ is the longest common affix string between the two words, and $Len(w)$ is a length of lemma w. Notice that the longest common affix is either common suffix or prefix between the two given lemmas e.g. 'compet' is the LCA between the lemmas 'compete' and 'competition'.

Table 7.1 shows some examples of lexical relations with their LCA and DM values in 5 languages. As can be seen, the DM values are vary due to language diversity in the derivational morphology. In a theory, the threshold β should be

Algorithm 2: Discovery of lexical relations in given languages

Input : R, a set of lexical relations from PWN, and L is the set of all languages

Output S, a set of new lexical relations

:

$S \leftarrow \emptyset$;
for *each* *lexical relation* $r \in R$ **do**
  for *each* *language* $l \in L$ **do**
    $con_1, con_2 \leftarrow Concepts(r)$;
    $syn_1 \leftarrow Synset(con_1, l)$;
    $syn_2 \leftarrow Synset(con_2, l)$;
    if $syn_1 \neq \emptyset$ *and* $syn_2 \neq \emptyset$ **then**
      for *each* $s_1 \in Senses(syn_1)$ **do**
        for *each* $s_2 \in Senses(syn_2)$ **do**
          $sim \leftarrow DM(s_1, s_2)$;
          if $sim > \beta$ **then**
            $p \leftarrow C(s_1, s_2, type(r))$;
            $S \leftarrow S \cup \{p\}$;
return S;

determined in each language. However, finding β manually for each language is very unreasonable because there is over 300 languages and the coverage of each language is unbalanced. In order to determine ideal value for β, we performed several grid searches for 36 languages, *having more than X words*. Since we wants to maximize high accuracy and reasonable recall, the threshold value of $\beta = 0.5$ is finally chosen.

7.3 Result

Table 7.3 shows the results for the 26 wordnets with than 600 lexical relations. Furthermore, there are 53 languages with more than 300 lexical relations. Overall our resource has now 186,048 lexical relations (only including Pertain, Antonym, Derived relations) in over 300 languages.

As can be seen from Table 7.3, the coverage of the discovered lexical relations is simultaneously improved while the number of senses for each language is being increased. Also, the coverage of pertain relation depends critically on the number of adjective senses.

7.4 Evaluation

We then evaluated samples of the lexical relations in 8 languages. For each language, we chose at least 100 random lexical relations and asked the native speakers to validate them manually. Table 7.2 provides the distributions of the chosen random lexical relations with the accuracy, provided by the speakers. As can be seen, the average accuracy of 8 languages is found to be 95.2.

Table 7.2: Precision of Discovered Lexical Relations

Languages	**Derived**	**Pertain**	**Antonym**	**Correct**	**Total**	**%**
Chinese	33	34	36	100	103	97
Spanish	35	35	35	104	105	99
Italian	344	0	156	489	500	97.8
Hindi	43	11	11	51	55	92.3
Hungarian	44	35	21	90	100	90
German	35	35	35	-	105	-
Russian	35	35	35	-	105	-
Arabic	35	35	35	-	105	-
Average						95.2

During the evaluation, we noticed that the wiktionary-based languages have significantly lower performances due to the lower qualities than those wordnets, developed manually. In the fact, the wiktionary-based resources, we used here from Open Multilingual WordNet, have the 90% alignment accuracy.

Table 7.3: The languages in UKC (with more than 600 lexical relations)

ISO	Languages	Synsets	Senses	Noun	Adj.	Pertain	Anto.	Derived	Total
eng	English	109,942	191,523	129,611	31,302	7,465	4,017	35,507	46,989
fin	Finnish	107,989	172,755	115,114	28,935	3,155	424	13,439	17,018
jpn	Japanese	51,366	151,262	92,755	17,783	1,710	942	14,167	16,819
zho	Chinese	98,324	123,397	77,458	26,306	1,262	981	9,635	11,878
fra	French	53,588	90,522	60,708	10,760	1,601	411	9,266	11,278
ron	Romanian	52,716	80,001	51,385	8,183	389	724	9,436	10,549
slv	Slovene	40,233	67,866	41,020	12,320	691	507	6,937	8,135
msa	Malay	31,093	93,293	38,331	11,735	365	92	7,444	7,901
ind	Indonesian	31,541	92,390	37,060	11,126	340	69	7,218	7,627
cat	Catalan	42,256	66,357	47,127	7,672	108	413	5,398	5,919
por	Portuguese	38,609	60,530	42,177	8,336	675	340	4,807	5,822
pol	Polish	35,083	87,065	64,415	9,810	317	330	4,643	5,290
tha	Thailand	65,664	83,818	62,927	5,847	294	165	4,354	4,813
spa	Spanish	35,232	53,140	34,336	6,954	392	227	3,737	4,356
hrv	Croatian	21,302	45,929	25,110	2,581	262	680	3,253	4,195
eus	Basque	28,848	48,264	38,871	148	2	103	3,640	3,745
nld	Dutch	28,253	57,706	47,571	1,140	179	157	2,860	3,196
deu	German	18,418	28,147	20,295	3,735	234	93	1,892	2,219
slk	Slovak	15,808	37,988	23,591	1,605	192	225	1,698	2,115
rus	Russian	18,392	31,826	21,664	4,056	284	111	1,686	2,081
glg	Galician	15,790	23,344	15,174	6,756	36	738	507	1,281
arb	Arabic	9,576	20,745	13,659	761	110	17	595	722
ita	Italian	33,560	42,381	42,324	32	0	208	500	708
hun	Hungarian	9,255	11,926	8,570	1,517	68	21	608	697
sqi	Albanian	4,671	9,593	6,033	80	0	17	678	695

Noun shows the coverage of the total noun senses for the corresponding language.
Adjective shows the coverage of the total adjective senses for the corresponding language.

Chapter 8

Cognates

In the historical linguistics and comparative linguistics, the identification of cognate sets is a key problem because of at least following few reasons.

- The identified cognates explicitly express a culture and history between two languages of the cognates, and they are important source of information for other fields of history studies e.g. archaeology [Renfrew, 1990] and paleogenetics [Haak et al., 2015].
- Based on this evolutional evidence among cognates across languages, the phylogenetic tree of languages has been built by the historical linguistic experts [Jäger, 2018].
- The machine translation methods achieve a better performances by exploiting cognate databases [Kondrak et al., 2003, Karakanta et al., 2018, Aires et al., 2016, Sennrich et al., 2015]. However, the data used in such methods are relatively small with a comparison with the database we extracted in this study.

Over the past few decades, a large existing body of algorithmic methods have been proposed for automatic identification of cognate sets, as well as that several etymological databases across different phyla have been built to support this interesting research field. However, there are two issues to use or integrate

the state-of-the-art databases with multilingual lexical resource like Babelnet or OMW.

- First is a number of concepts covered by the databases is ranged between 50 and 200 concepts even though a highest number of languages is up to 4000.
- Second is that a word form used by the databases is a phonetic transcriptions while this form is simply very hard to integrate with the word scripts used by the multilingual lexical resources.

In this chapter, we propose a fully-automatic method to extract cognate sets from the existing multilingual resources. The main technical contributions are as follows:

- Lingtra, Multilingual transliteration tool
- Algorithm to detect cognates
- Empirical dataset on evaluation

The remainder of the chapter is as follows. Section 8.1 describes the method. Section 8.2 presents an evaluation of the method. Section **??** provides the analyses and findings about how internationalism are globalizing a world, and how it is revealed by the detected cognates. Section 8.4 concludes this chapter.

8.1 Method

The main task in this chapter is to detect cognates in UKC by harvesting a possible information and tools. In the subsection 8.2.1, we first introduces Lingtra that is Multilingual Transliteration tool. The subsection 8.2.2 describes Etymological WordNet (EWN). Finally, we provides an algorithm to detect cognates from UKC by using Lingtra and EWN.

8.1.1 Lingtra - Mutlilingual Transliteration tool

The resource we use has 335 languages, covered by 40 scripts e.g. Arabic, Devanagri, Kanji, Cyrillic, and more. In this subsection, I introduce the multilingual transliteration tool called Lingtra. The main purpose of Lingtra is to transliterate a unicode text to latin alphabets which is also called Romanization. Lingtra is a python tool that employees the international standard rules and codes, developed by the Wiktionary community (the largest community of lexicography).

Languages	**Word**	**Expected**	**Junidecode**	**Google**	**WikTra**
English	book	book	book	book	book
Malayalam	മലയാളം	malayāḷaṃ	mlyaallN	malayāḷaṁ	malayāḷaṃ
Arabic	نواة	nawātun	nw@	nawa	nawātun
Japanese	コンピュータ	konpyūtā	konpiyuta	konpyūtā	konpyūtā
Thai	ราชาทิราด	rā chā tí râat	raachaathiraad	rā chā thi rād	raa-chaa-tí-râat
Russian	москва	moskva	moskva	moskva	moskva
Hindi	देवनागरी	devnāgrī	devnaagrii	devanaagaree	devnāgrī
Bengali	বাংলা	bangla	baaNlaa	bānlā	bangla
Greek	ναῦτέω	anaūtéō	anauteo	anāftéo	anaūtéō
Kashmiri	کَمپیوٗٹَر	kampeūṭar	khampy[?]w?ttar	-	kampeūṭar
Persian	پیتزا	pitzâ	pytz	-	pitzâ
Hebrew	יששכר	yiśśāḵār	yshshkr	yissachar	yiśśāḵār
Tamil	ரெஃஸ்	rex	reHs	reḥs	rex
Ethiopic	አዲስ አበባ	ʾädis ʾäbäba	ʿaadise ʾaababaa	ādīsi ābeba	ʾädis ʾäbäba
Bodo	ཁ་པར	kha par	kh-pr	-	kha par

Lingtra in Japanese language only work with scripts of Hiragana and Katakana.
Lingtra in Thai language only work with a sequence of syllables.

Table 8.1: Comparison with the state-of-the art transliteration tools

I have compared the results of Lingtra with the two state-of-the-art transliteration tools, namely: Junidecode and Google.

- *Junidecode*[1] is a Java port of Text::Unidecode perl module by Sean M.

[1]Junidecode is publicly available at https://github.com/gcardone/junidecode. Accessed on 13.10.18

Burke [Burke, 2001] that takes Unicode data and tries to represent it in US-ASCII characters. The author of Unidecode first manually created a big table of unicode characters with US-ASCII characters. And this table is exploited to transliterate each character of a given word to US-ASCII character.

- Google Transliteration is a dictionary based phonetic transliteration approach, and is not publicly available. A small amount of money is needed to use the service automatically. In this comparison, we collected results of the words in table 8.1 by manually checking from Google Translate[2] online.

Table 8.1 represents a comparison between expected words and the transliterated words by Junidecode, Google Transliteration, and Lingtra. From this table, we made the following observations:

- Junidecode offers a wide range of transliteration across all languages while its expected quality is relatively erroneous than Google and Lingtra methods.
- Google offers a high quality while its language support are limited to smaller number of languages than Lingtra.
- Lingtra provides more accurate words in the comparison table than Google and Junidecode.

In overall, Lingtra offers a wide range of transliterations over 40 different scripts across more than 200 languages. However, it has a small number of issues with a few languages including Thai and Japanese. In Thai, the Lingtra can be only used for monosyllabic words, and for multisyllabic words an additional tool is needed to recognize syllables of the words. Then every syllable of the word can be transliterated by Lingtra. In Japanese, Lingtra only works with scripts

[2]https://translate.google.com/

of Hiragana and Katakana but not with Kanji (Chinese characters). Therefore, in order to improve a quality of our transliterated words, we have used a few monolingual transliteration tools in those languages e.g. Kuromoji[3] in Japanese language.

8.1.2 Etymological WordNet

To improve the performance and results of our method, we decided to use Etymological WordNet[4] (EWN) [De Melo, 2014] that is a lexical resource covering several types of lexical relations between words including derivational relations and etymology relations. EWN was automatically built by harvesting etymological information encoded in Wiktionary. In this work, we have only used its cross-lingual etymological relations as counted 94,832 relations.

8.1.3 Algorithm

Our goal is to extract a set of cognate instances between words of a given concept of interest. The algorithm we propose is based on the following intuitions:

- if two words are orthographically aligned, then they are cognates. In this level of alignments, the decision is made on only texts of original scripts.
- if two words are orthographically similar in its translitrated words, then they are cognates. In this case, the decision is made on both levels of text: the original scripts and the transliterated words by Lingtra.
- if two words are etymologically related, then they are cognates. In this case, the lexical resource of Etymological WordNet is exploited to find if they has same source of etymologies in history.

[3]https://github.com/atilika/kuromoji. Accessed on 13.10.18

[4]EWN data is available at http://www1.icsi.berkeley.edu/ demelo/etymwn/. Accessed on 10.14.18

Based on all above information, we built a undirected graph where each node represents a word and each edge between two nodes represents a cognate relation between two words. Then our intuition is that each connected subgraph represent one group of cognate words. A single node that is not connected any other node of the graph is called a isolated word. The algorithm is structured as follows.

Algorithm 3: A Cognate Discovery Algorithm

Input : c, a lexical concept
Input : $\mathcal{R}$, a multilingual lexical resource
Output : S, a set of cognates for the concept c.

1 $V, E, S \leftarrow \emptyset$;
2 $\mathcal{L} \leftarrow \text{Languages}_{\mathcal{R}}(c)$;
3 **for** *each language* $l \in \mathcal{L}$ **do**
4 **for** *each word* $w \in \textit{Words}_{\mathcal{R}}(c_1, l)$ **do**
5 $V \leftarrow V \cup \{\, v = < w, l > \,\}$;
6 **for** *each node* $v_1 = <w_1, l_1> \in V$ **do**
7 **for** *each node* $v_2 = <w_2, l_2> \in V$ **do**
8 **if** $l_1 \neq l_2$ **then**
9 **if** $w_1 = w_2$ ***or*** $\textit{orthoSim}(w_1, w_2)$ **then**
10 $E \leftarrow E \cup \{\, e = < v_1, v_2 > \,\}$;
11 **else if** $\textit{orthoSim}(\textit{Lingtra}(w_1, l_1), \textit{Lingtra}(w_2, l_2))$ **then**
12 $E \leftarrow E \cup \{\, e = < v_1, v_2 > \,\}$;
13 **else if** $\varphi(w_1, l_1) \bigcap \varphi(w_2, l_2) \neq \emptyset$ **then**
14 $E \leftarrow E \cup \{\, e = < v_1, v_2 > \,\}$;
15 $G \leftarrow < V,\ E >$;
16 **for** *each node* $v_1 = <w_1, l_1> \in V$ **do**
17 **for** *each node* $v_2 = <w_2, l_2> \in V$ **do**
18 **if** $\gamma_G(v_1, v_2) = true$ **then**
19 $S \leftarrow S \cup \{\, \text{x} = <w_1, l_1, w_2, l_2>\}$;
20 **return** S;

Step 1. (Lines 1-2). In Line 1, it initializes the variables V, E, S by an empty set where the variable V is intended to store a set of nodes of the graph G while E is intended to store a set of edges of the graph G. In line 2 it collects in $\mathcal{L}$ all

Algorithm 4: A Cognate Discovery Algorithm

Input : c, a lexical concept
Input : $\mathcal{R}$, a multilingual lexical resource
Output : S, a set of cognates for the concept c.

1 $V, E, S \leftarrow \emptyset$;
2 $\mathcal{L} \leftarrow \text{Languages}_{\mathcal{R}}(c)$;
3 **for** *each language* $l \in \mathcal{L}$ **do**
4 **for** *each word* $w \in \textit{Words}_{\mathcal{R}}(c_1, l)$ **do**
5 $V \leftarrow V \cup \{\, v = <w, l> \,\}$;
6 **for** *each node* $v_1 = <w_1, l_1> \in V$ **do**
7 **for** *each node* $v_2 = <w_2, l_2> \in V$ **do**
8 **if** $l_1 \neq l_2$ **then**
9 **if** $w_1 = w_2$ ***or*** $\textit{sim}(w_1, w_2)$ **then**
10 $E \leftarrow E \cup \{\, e = <v_1, v_2> \,\}$;
11 **else if** $\textit{sim}(\textit{WikTra}(w_1, l_1), \textit{WikTra}(w_2, l_2))$ **then**
12 $E \leftarrow E \cup \{\, e = <v_1, v_2> \,\}$;
13 **else if** $\varphi(w_1, l_1) \bigcap \varphi(w_2, l_2) \neq \emptyset$ **then**
14 $E \leftarrow E \cup \{\, e = <v_1, v_2> \,\}$;
15 $G \leftarrow <V, E>$;
16 **for** *each node* $v_1 = <w_1, l_1> \in V$ **do**
17 **for** *each node* $v_2 = <w_2, l_2> \in V$ **do**
18 **if** $\gamma_G(v_1, v_2) = true$ **then**
19 $S \leftarrow S \cup \{$ x = $<w_1, l_1, w_2, l_2>\}$;
20 **return** S;

the languages where the given concept c is lexicalized;

Step 2. (Lines 3-5). It collects all nodes of G into V by considering a word w as a individual node if a word w in a language l expresses c.

Step 3. (Lines 6-8). It iterates every possible pair of nodes in V (Lines 7-8). In the following steps, if the pair of nodes exhibits a cognate, it adds an edge between the pair of nodes into E. (Line 9) It skips a pair if two nodes have same language.

Step 3. (Lines 9-10). It tries to recognize if w_1 and w_2 are cognates. This attempt succeeds if one of two conditions hold: (i) the two words are same, or (ii) the two words are *orthographically* similar, as computed by the function morphSim. If it succeeds it adds $e =<v_1, v_2>$ to the set of edges E.

$$\text{sim}_a(w_1, w_2) = \frac{2 * \text{len}(\text{LCS}(w_1, w_2))}{\text{len}(w_1) + \text{len}(w_2)} \tag{8.1}$$

$$\text{dis}(w_1, w_2) = 1.0 - \text{Min}(\frac{\text{dis}_a(lan(w_1), lan(w_2))}{T_D}; 1.0) \tag{8.2}$$

$$\text{sim}(w_1, w_2) = \text{sim}_a(w_1, w_2) + T_I * \text{dis}(w_1, w_2) > T_S \tag{8.3}$$

Step 4. (Lines 11-12). In this step, Lingtra is used to transliterate words in the corresponding languages (Line 11). Then it checks if the resulted words are cognates *orthographically*. If it succeeds it adds $e =<v_1, v_2>$ to the set of edges E (Line 12).

Step 5. (Lines 13-14). In this step it exploits EWN to recognize if the two words are cognates. The function φ returns a set of ancestors words of the given word and language in EWN. This attempt succeeds if the two words have a least one common ancestor word (Line 13). If it succeeds it adds $e =<v_1, v_2>$ to the set of edges E (Line 14).

Step 6. (Lines 15-20). In last step, a set of cognate instances is detected from the graph G=<V, E>. It iterates every possible pair of nodes v_1, v_2 from V (Lines 16-17). Then for the given pair the function γ_G checks if the two nodes are connected in the graph G (Line 18). If yes, it creates a cognate instance x as <w_1, l_1, w_2, l_2>, and adds it into S, the set of cognate instances (Line 19). Finally, it returns S, a set of cognate instances for the concept c.

8.2 Evaluation and Results

We organize this section in three parts. First, we describe how the evaluation dataset is built. Second we provide the detailed information of how the parameters are tuned with the dataset. Third, we analyse the results of experiment and evaluate it again.

8.2.1 Dataset Annotation

In this experiment, we first wanted to use the existing cognate dataset to this evaluation. However, most of the existing cognate databases with different phyla are encoded in forms of phonetic transcriptions, often in International Phonetic Alphabet (IPA), and have no words in its original scripts of their languages.

Therefore, we created a dataset of 50 concepts with the fully annotated sets of cognate groups. Those concepts ranges from 22 words to 163 words while its languages ranges from 16 to 133. For each concept, we asked two linguistic experts to find different groups of cognates among words of the given concept. The experts made the decisions based on the online resources like Wiktionary and Online Etymology Dictionary[5].

8.2.2 Algorithm configuration

In this task, the only hyperparameter to be learned is the orthographic threshold T_M for the function 8.2. We have selected a grid of $T_M = [1.0; 2.0]$ (). In this grid, we computed the ideal value for T_M with an increment of 0.01 by achieving the best performance on the dataset provided in the previous subsection 8.2.1. With this settings, our method in Algorithm 4 with Lingtra and EWN achieved significantly good performances of Adjusted Rand Index (ARI) [Fowlkes and Mallows, 1983] as can be seen from Table 8.2.

[5]https://www.etymonline.com/. Accessed on 14.10.18

Table 8.2: Parameter configuration and comparisons.

Methods	ARI for sample concepts						ARI
	computer	apple	snake	song	lion	kungfu	
Baseline[1]	0.484	0.525	0.656	0.621	0.401	0.529	0.564
Our method+EWN[2]	0.674	0.571	0.685	0.642	0.685	0.544	0.632
Our method+Lingtra[3]	0.731	0.867	0.835	0.827	0.694	0.949	0.820
Our method+Lingtra+EWN[4]	0.928	0.899	0.849	0.835	0.823	1.000	0.888

[1] Skips the Lines 11-14 in the Algorithm 4

[2] Skips the Lines 11-12 in the Algorithm 4

[3] Skips the Lines 13-14 in the Algorithm 4 and $T_M = 1.71$

[4] Our method + Lingtra + EWN: $T_M = 1.71$

The baseline method achieves relatively poor results from others. The third method with only Lingtra perform significantly better than other methods without Lingtra. And our performance is significantly improved when both combination of Lingtra and EWN. Table 8.2 also shows that ARI for six sample concepts of *computer*, *apple*, *snake*, *song*, *lion*, *kungfu*. Figure 8.1 shows the resulted cognate sets for the sample concept *song*. This result is also even browsable at Linguarena online[6].

By analysing the generated cognate sets for the concept *song*, we can observe some very interesting facts as follows.

- By learning only five words representing the first five cognate groups, we can speak in a majority of all the UKC languages.
- The first group (canzone - red) belongs to the romance subphylum while the second group (songu - blue) belongs to the germanic subphylum
- The most diverse part of the world is a west side of the black sea.
- etc...

These kinds of facts could be very useful to the historical linguists.

[6] http://linguarena.eu/view/763193744/en

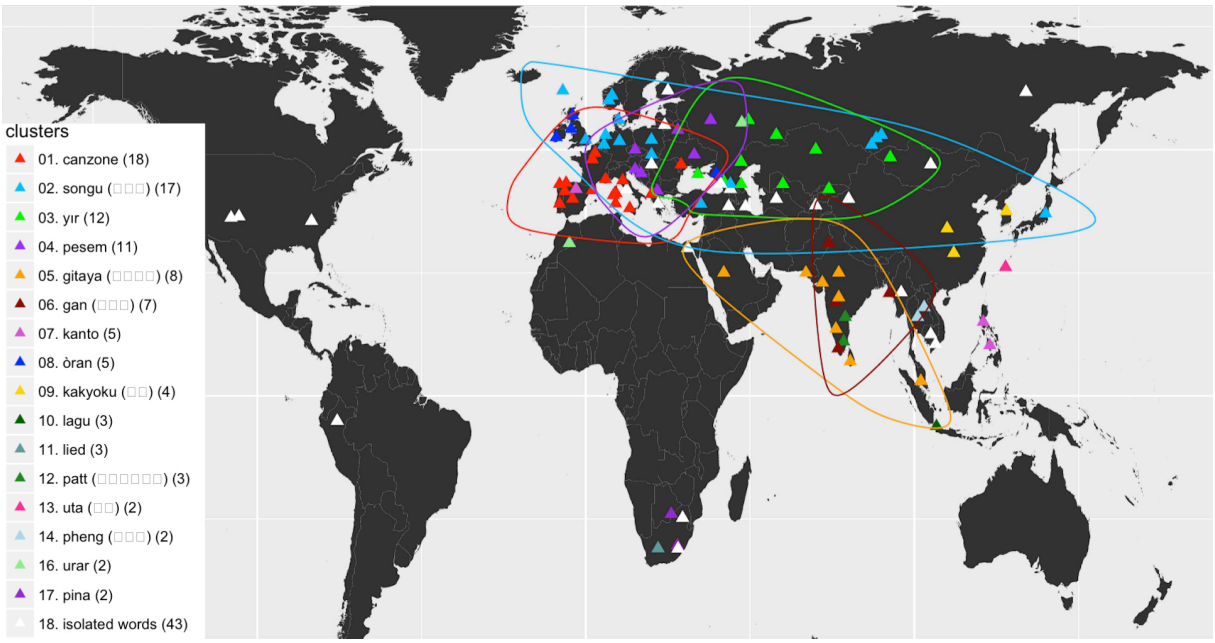

Figure 8.1: The generated cognate sets of a concept 'song'

8.2.3 Results

Table 8.3: Cognate Groups.

Group	continent	phylum	#Cognates	by %	Examples
a'	same	same	1,027,960	37.2%	ita:[cultura] gla:[cultar]
b'	different	same	110,866	4.1%	eng:[doctor] ben:[ডাক্তার (daktar)]
c'	same	different	936,734	33.8%	eus:[ministro] rus:[министр (ministr)]
d'	different	different	688,972	24.9%	swe:[graf] mal:[ഗ്രാഫ് (graph)]
Total			2,764,532	100.0%	

As results of Algorithm 4, the UKC nows contains 2,764,532 cognate instances across its pool of 335 languages. We grouped the cognates into four groups with different values of attributes of phylum and continent. The distribution of the groups and its example are shown in Table 8.3.

Observation 1. The table shows that the vast majority of total cognates (71.06 % as $a' + c'$) are exhibited within same continent.

Observation 2. The group b', the cognate group with different continent and same phylum, is a relatively fewer than any other groups.

Table 8.4: Cognate accuracies for the samples

Group	Samples	Accuracy%	Confidence	F-Accuracy%
a'	100	97	4.97	96.42
b'	100	93	4.88	90.77
c'	100	92	4.98	91.63
d'	100	94	4.87	91.56
Total	400	94	4.93	92.59

We then evaluated cognate samples of the four groups. For each group, we randomly chose 100 cognate instances and asked the linguistic experts to validate them manually with their confidence scores. They were provided with the two words w_1 and w_2 in the languages l_1 and l_2 respectively. And they were asked the following question: "Do you think words w_1 in the language l_1 and w_2 in language l_2 are cognates?", and they had to provide a yes/no answer with their self confidence score which ranges between 1 to 5 (the value higher means that answer is very confident while lower value represent lesser confidence).

Table 8.4 provides the accuracies and confidence scores for the four groups, provided by the evaluators. As can be seen, the average accuracy of four groups is found to be 94% with their average confidence 4.93.

8.3 Impacts of Internationalism

In linguistics, an internationalism or international word is a loanword that occurs in several languages (that is, translingually) with the same or at least similar meaning and etymology. In this section, we aim to investigate a relationship between cognates and internationalism. First, we start to describe how the internationlism of a concept can be measured. Second, we analyse evidently how in-

ternationalism of languages and world reflects to a lexical space of words across languages.

8.3.1 Quantifying Concept Internationalism

The idea behind of a computation of *Concept Internationalism*, IntMea, is that if a concept c exhibits an international word then most of the UKC words, denoting a concept c should be a cognate with one another. On this measure, we proposed the following formula:

$$\text{CogClaLan}_c(w) = \bigcup_{x \in \text{Cognates}(c)} \{\text{Languages}(x) \mid \text{If words}(x) \cap w \neq \emptyset\} \tag{8.4}$$

$$\text{IntMea}(c) = \frac{(|\text{Languages}(c)| - 1) * \max_{w \in \text{words}(c)}(|\text{CogClaLan}_c(w)|)}{|\text{Languages}(c)|^2} \tag{8.5}$$

where the function *Cognate Class Languages*, $\text{CogClaLan}_c(w)$, returns a set of languages where a concept c is denoted as cognate word by given word w, Concept Internationalism Measure, $\text{IntMea}(c)$ returns a value between 0 and 1. If it is closer to 1, it states that a given concept c exhibits internationalism across a world (e.g. pizza, kung-fu, tennis). If it is closer to 0, it states the concept c is a equally fundamental and natural to all languages (e.g. fish, tree, axe).

In other theoretical words, when physical encounters of the concept c are first time experienced by speakers of a new language, that language community had no similar or close physical experiences and often forced to adopt a foreign word to their language.

8.3.2 Analyses of Internationalism

Figure 8.2 displays the median distributions of internationalism measures over 5 different groups of all 109,942 concepts with different settings of concept coverages. As can be seen from here, when a concept coverage is a smaller or equal to 10 languages, group's median value is at 0.122 of internationalism while the concept's coverage increases than 10 languages or 20 languages then its median value converges to 0.25 of internationalism. From here, we made the following observations:

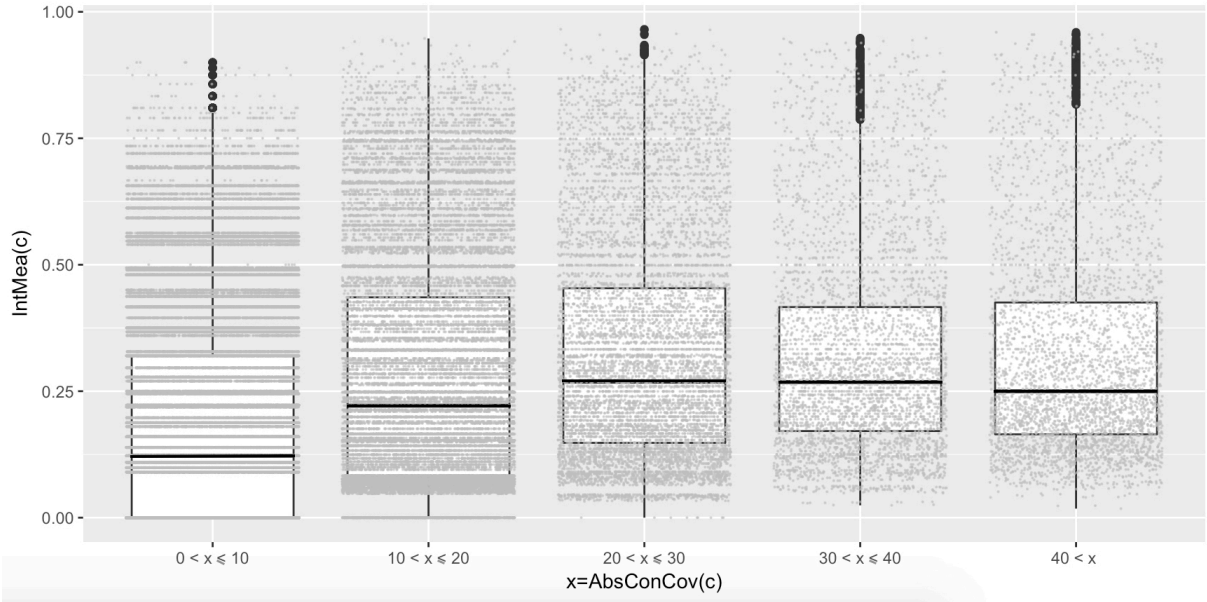

Figure 8.2: Median Internationalism Measures to Concepts of UKC

Observation 3. The concepts, exhibiting internationalism, are treated as outliers to a rest of all concepts.

Observation 4. The concept internationalism measure is very consistent to all different levels of concept coverage settings.

Figure 8.3 shows the distributions of all 2.7 million cognates over two dimensions of geographic distance (1 in 1000 km) and internationalism measure. For

each cognate x, we estimated two values as follows.

- The internationalism value of a cognate x is same as an internationalism value of its concept c_x. So Intmea(x) = Intmea(c_x).
- Second value, the geographic distance, is estimated as a distance between two locations of its two languages l_1 and l_2 in the given cognate x. In UKC, every language has its location as given in a form of latitude and longitude.

8.4 Summary

In this chapter, we introduced Lingtra, Multilingual Transliterations Tool, and its comparison with the state-of-the-art systems. From this comparison, it was crystal that Lingtra outperforms other existing systems in the two terms of measures. The first measure is that Lingtra offers much broader range of scripts to be transliterated (more precisely 40 different scripts). The second measure is that Lingtra provides more high quality than Google and Unidecode as shown in Table 8.1.

We presented a two-fold evaluation approach to validate if the results is a high quality. In the results section, we pointed out very interesting natures of cognates, previously unrevealed empirically, as follows.

- The vast majority of cognates (72% percents) happens in same continent. This is actually easy to guess for linguistic experts. The interesting one is next.
- Among the cognates within same continent, the amount of different-phylum cognates are almost equal to the amount of same-phylum cognates. In this new basis, we can think that cultures are passed through words across genetically unrelated languages as same amount as related languages.

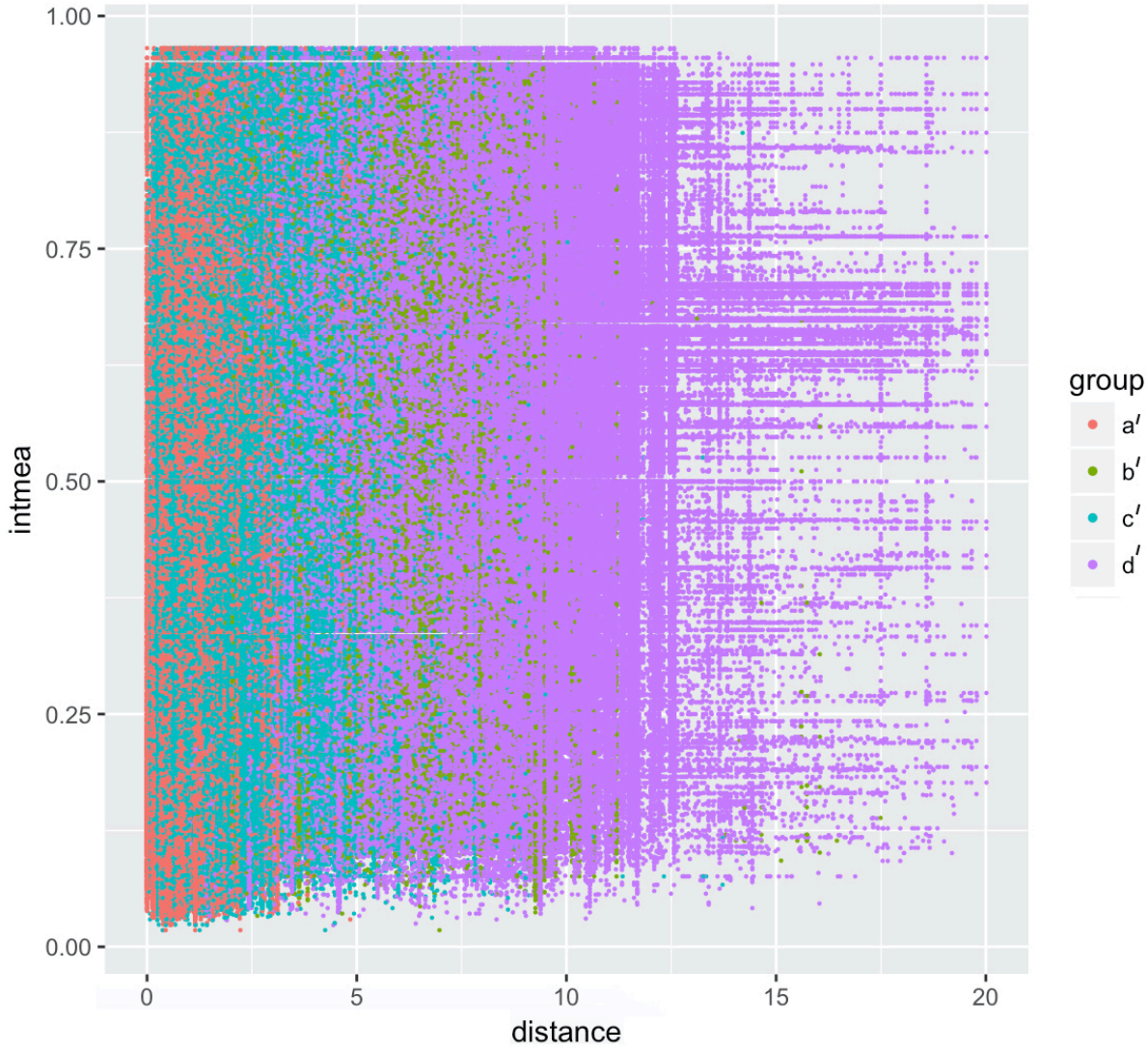

Figure 8.3: Distributions of 4 different groups of cognates over two dimensions of geographic distance (1 in 1000 km) and internationalism measure.

Chapter 9

Conclusion and Future Work

> "I do the very best I know how - the very best I can; and I mean to keep on doing so until the end."
> —-*Abraham Lincoln*

In this thesis, in order to solve the fundamental linguistic problems of lexical semantics, we proposed several sets of a quantitative, formal measures to language diversity, resource incompleteness, and resource incorrectness. By exploiting those measures, we have showed that how linguistic problems like distinguishing homonyms, and finding cognates could be solved in a precise,diversity-aware manner. As results of these case studies, we have enriched UKC a lot at the linguistic levels, and is even browseable online at the website of Linguarena[1].

In our future study, we will consider to solve other fundamental linguistic problems (such as implicational universals of lexical elements) and enrich UKC, and also try to expand and improve the quantitative measures. Future work also concerns deeper analyses and investigation of the hypotheses arised from the experiments of this thesis.

[1] http://linguarena.eu/

Bibliography

José Aires, Gabriel Lopes, and Luís Gomes. English-portuguese biomedical translation task using a genuine phrase-based statistical machine translation approach. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 456–462, 2016.

Ju D Apresjan. Regular polysemy. *Linguistics*, 12(142):5–32, 1974.

Mark Aronoff and Janie Rees-Miller. *The handbook of linguistics*, volume 43. John Wiley & Sons, 2003.

Alan Bell. Language samples. universals of human language, ed. by joseph greenberg et al., 1.153-202, 1978.

Gabor Bella, Fausto Giunchiglia, and Fiona McNeill. Language and domain aware lightweight ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2017.

Luisa Bentivogli and Emanuele Pianta. Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*, pages 8–12. Stuttgart: Universität Stuttgart, 2000.

Pushpak Bhattacharyya. Indowordnet. In *The WordNet in Indian Languages*, pages 1–18. Springer, 2017.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. Introducing the arabic wordnet project.

In *Proceedings of the third international WordNet conference*, pages 295–300. Citeseer, 2006.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362, 2013.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016, 2016.

Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

Paul Buitelaar. *CoreLex: systematic polysemy and underspecification*. PhD thesis, Citeseer, 1998.

Sean M Burke. Unidecode! *Sys Admin*, 10(12):54–60, 2001.

William Croft. *Typology and universals*. Cambridge University Press, 2002.

David Crystal. *The Cambridge encyclopedia of the English language*. Ernst Klett Sprachen, 2004.

Jurgita Cvilikaitė. Lexical gaps: resolution by functionally complete units of translation. *Darbai ir dienos, 2006, nr. 45, p. 127-142*, 2006.

Gerard De Melo. Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154. Citeseer, 2014.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Jia Deng, Olga Russakovsky, Jonathan Krause, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Scalable multi-label annotation. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2014.

Nicholas Evans and Stephen C Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(05):429–448, 2009.

Nicholas Evans and Hans-Jürgen Sasse. *Problems of polysynthesis*, volume 4. Oldenbourg Verlag, 2002.

Christiane Fellbaum, Anne Osherson, and Peter E Clark. Putting semantics into wordnet' s" morphosemantic" links. In *Language and Technology Conference*, pages 350–358. Springer, 2007.

Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383): 553–569, 1983.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. A taxonomic classification of wordnet polysemy types. In *Proceedings of the 8th GWC Global WordNet Conference*, 2016.

Radovan Garabík and Indrė Pileckytė. From multilingual dictionary to lithuanian wordnet. *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80, 2013.

Fausto Giunchiglia and Mattia Fumagalli. Concepts as (recognition) abilities. In *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, volume 283, page 153. IOS Press, 2016.

Fausto Giunchiglia and Mattia Fumagalli. Teleologies: objects, actions and functions. In *Proceedings of the 36th International Conference on Conceptual Modeling (ER 2017)*, 2017.

Fausto Giunchiglia, Mladjan Jovanovic, Me rcedes Huertas-Migueláñez, and Khuyagbaatar Batsuren. Crowdsourcing a Large-Scale Multilingual Lexico-

Semantic Resource. In *The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15), San Diego, CA*, 2015.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. Understanding and exploiting language diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4009–4017, 2017.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529, 2012.

Joseph H Greenberg. Universals of language. 1966.

Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, et al. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207, 2015.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 2.7. *Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany). Available at glottolog. org/. Accessed February*, 27:2017, 2015.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, June 2010. ISSN 1469-1825. URL `http://journals.cambridge.org/abstract_S0140525X0999152X`.

Shu-Kai Hsieh and Yu-Yun Chang. Leveraging morpho-semantics for the discovery of relations in chinese wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 283–289, 2014.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. Chinese wordnet: Design, implementa-

tion, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23, 2010.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *CoRR*, abs/1802.06079, 2018. URL `http://arxiv.org/abs/1802.06079`.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, pages 1–23, 2018.

Göran Kjellmer. Lexical gaps. *Language and Computers*, 48(1):149–158, 2003.

Svetla Koeva. Derivational and morphosemantic relations in bulgarian wordnet. *Intelligent Information Systems*, 16:359–369, 2008.

Svetla Koeva, Stoyan Mihov, and Tinko Tinchev. Bulgarian wordnet–structure and validation. *Romanian Journal of Information Science and Technology*, 7 (1-2):61–78, 2004.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 46–48. Association for Computational Linguistics, 2003.

Adrienne Lehrer. Notes on lexical gaps. *Journal of Linguistics*, 6(2):257–261, 1970.

Krister Lindén and Jyrki Niemi. Is it possible to create a very large wordnet in 100 days? an evaluation. *Language resources and evaluation*, 48(2):191–201, 2014.

Krister Lindén, Jyrki Niemi, and Mirka Hyvärinen. Extending and updating the finnish wordnet. In *Shall We Play the Festschrift Game?*, pages 67–98. Springer, 2012.

John Lyons. Semantics. *Cambridge University Press, London, England*, 1977.

Michael Matuschek, Christian M Meyer, and Iryna Gurevych. Multilingual knowledge in aligned wiktionary and omegawiki for translation applications. *Translation: Computation, Corpora, Cognition–Special Issue on ' Language Technology for a Multilingual Europe*, 3(1):87–118, 2013.

April McMahon and Robert McMahon. *Language classification by numbers*. Oxford University Press on Demand, 2005.

George A Miller. Nouns in wordnet: a lexical inheritance system. *International journal of Lexicography*, 3(4):245–264, 1990.

Ruth Garrett Millikan. *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press, 2000.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics, 2010.

Karel Pala and Dana Hlaváčková. Derivational relations in czech wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 75–81. Association for Computational Linguistics, 2007.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299, 2009.

Wim Peters. Metonymy as a cross-lingual phenomenon. In *Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14*, pages 1–9. Association for Computational Linguistics, 2003.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Developing an aligned multilingual database. In *Proc. 1st Int ' l Conference on Global WordNet*. Citeseer, 2002a.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Multi-wordnet: developing an aligned multilingual database.". In *Proceedings of the First International Conference on Global WordNet, Mysore, India, January*, pages 21–25, 2002b.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45 (2):121–142, 2011.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. Open dutch wordnet. In *Global WordNet Conference*, page 300, 2016.

Colin Renfrew. *Archaeology and language: the puzzle of Indo-European origins*. CUP Archive, 1990.

Jan Rijkhoff, Dik Bakker, Kees Hengeveld, and Peter Kahrel. A method of language sampling. *Studies in Language. International Journal sponsored by the Foundation " Foundations of Language"* , 17(1):169–203, 1993.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Gary F Simons. *Ethnologue: Languages of the world*. sil International, 2017.

Mahesh Srinivasan and Hugh Rabagliati. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152, 2015.

Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137, 1955.

Morris Swadesh. *The origin and diversification of language*. Transaction Publishers, 1971.

Dan Tufis, Dan Cristea, and Sofia Stamou. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43, 2004.

Piek Vossen. Introduction to eurowordnet. *Computers and the Humanities*, 32 (2-3):73–89, 1998.

Piek Vossen, Wim Peters, and Julio Gonzalo. Towards a Universal Index of Meaning. In *SIGLEX99: Standardizing Lexical Resources*, pages 81–90, 1999. URL `http://www.aclweb.org/anthology/W99-0512`.

Piek Vossen, Francis Bond, and J McCrae. Toward a truly multilingual globalwordnet grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29, 2016.

Søren Wichmann, André Müller, Viveka Velupillai, Cecil H Brown, Eric W Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, et al. The asjp database (version 13). *URL: http://email. eva. mpg. de/˜ wichmann/ASJPHomePage. htm*, 3, 2010.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771, 2016.

Holly Young. The digital language divide. In *URL=http://labs.theguardian.com/digital-language-divide/*, 2015.