

# One World - Seven Thousand Languages

Fausto Giunchiglia, Khuyagbaatar Batsuren, Abed Alhakim Freihat

University of Trento, Trento TN 38100, Italy,  
fausto.giunchiglia@unitn.it

**Abstract.** We present a large scale multilingual lexical resource, the *Universal Knowledge Core* (UKC), which is organized like a *Wordnet* with, however, a major design difference. In the UKC the meaning of words is represented not only with *synsets* but also using language independent *concepts* which cluster together the synsets which, in different languages, codify the same meaning. In the UKC, it is concepts and not synsets, as it is the case in the Wordnets, which are connected in a semantic network. The use of language independent concepts allows for the native integrability, analysis and use of any number of languages, with important applications in, e.g., multilingual language processing, reasoning (as needed, for instance, in data and knowledge integration) and image understanding.

## 1 Introduction

Since the seminal work on the *Princeton WordNet* (PWN) [1] a lot of effort has been devoted to the production of large scale lexical resources, some of which are multilingual, see, e.g., [2–4].<sup>1</sup> These resources, also because largely constructed by “replicating” the PWN, share its basic organizational principles and, in particular, the fact that the multiple meanings of a *word* are codified as a *lexicalized concept*, also called a *synset*, consisting of a (possibly incomplete) set of synonymous words. Furthermore, in a multilingual Wordnet, each language is developed as if it was the only language and then, if  $s_1$  is the synset of a word  $w_1$  in a reference language  $L_1$ , usually PWN English, then the synset  $s_2$  of the corresponding word  $w_2$  in  $L_2$  (namely, the translation of  $w_1$  in  $L_2$ ) is linked to  $s_1$ .

Our goal in this paper is to introduce a multilingual lexical resource that we call the *Universal Knowledge Core* (UKC).<sup>2</sup> 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). The CC provides

---

<sup>1</sup> See <http://globalwordnet.org/> for a compilation of the most relevant resources available today.

<sup>2</sup> The word *knowledge* in *UKC* is motivated by our focus on studying language not *per se* but as a key component of reasoning systems.

**Table 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.

a uniform view over languages, it allows to compare them, to study their differences and similarities and to exploit this information to *measure* and *improve* the quality of linguistic resources and the UKC in particular.

The *diversity among languages* has been extensively studied in the fields of *Historical* and *Comparative Linguistics* [5] with, however, major limitations due to the problem that the data sets are very small (in the order of tens of elements). The work described in [6] provides a language diversity aware algorithm which can distinguish between homographs and polysemes. It is a first example of how the UKC paves the way to large scale quantitative studies of language diversity also, but not only, towards the production of high quality language resources.

In turn, the existence of the CC makes the UKC *not biased by any language and culture* and, therefore, *inherently open* and easily extensible. In particular, *lexical gaps*, namely missing concepts lexicalized in a new language can be dealt with by adding a new concept, thus solving one of the difficulties 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 [7]. 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., [6]. 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.<sup>3</sup>

Finally, it is 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 [8], 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 of the work in [9, 10] which uses Wordnet for the large scale classification

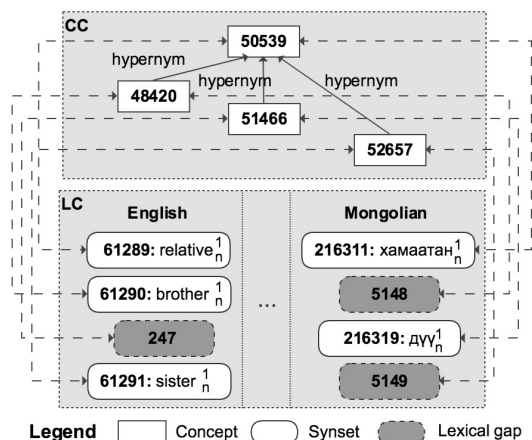
<sup>3</sup> From February 2018, the UKC will be browsable on line at the link <http://kidf.eu>.

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).

The paper is organized as follows. In section 2 we describe the organization of the UKC. In Section 3 we describe the complementary roles of words, synsets and concepts. In Section 4 we describe the resulting three layer organization of the UKC. In Section 5 we define the notion of language diversity. In Section 6 we deal with the issue of the quality of the resources and of the UKC in particular. The related work and our plans for the future work conclude the paper.

## 2 The Language and the 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*).



**Fig. 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.

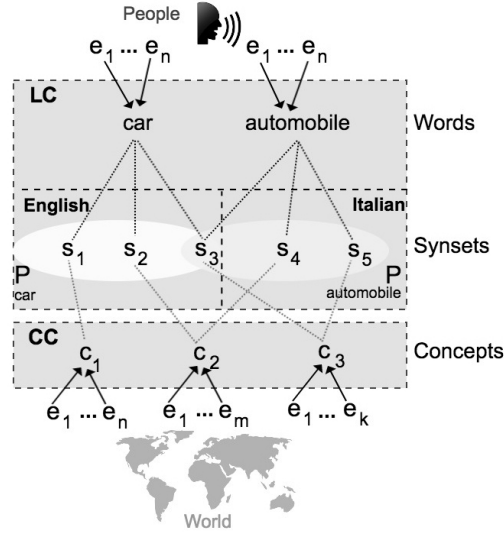
One difference with the PWN is that, in the UKC, lexical gaps have glosses, even if they do not have examples (which would be impossible). The intuition is that the gloss of a lexical gap can be seen as “local” language dependent description of a missing synset. This choice has turned out to be pragmatically useful when one is interested in understanding the meaning of a lexical gap without knowing the language(s) which generate(s) them.

### 3 Words, Synsets and Concepts

Humans build representations of what they perceive, what we usually call *the world*, as complex combinations of *concepts* where, following [11], 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,  $e_1, \dots, e_n$ , during which *substances manifest* themselves to a perceiver (e.g., an observer or a listener), where substances have two fundamental properties: (i) *they maintain some level of, but not full, invariance on how they manifest themselves to observers across multiple encounters* and (ii) *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 [11, 12] for a detailed discussion about these notions and also [13] for the early work in the field of *Biosemanantics* 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  $e_1, \dots, e_n$  during which they are perceived by, e.g., a listener or 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.



**Fig. 2.** The UKC and the World.

We represent words, synsets and concepts and their respective roles as in Figure 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  $c_1, \dots, c_n$ , while, at the top, there are words  $w_1, \dots, w_n$  (in Figure 2, *car* and *automobile*), where both words and concepts are perceived as the result of the encounters  $e_1, \dots, e_n$ .

Moving to the center of Figure 2, the synsets  $s_1, \dots, s_n$  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 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., Portuguese and Brazilian Portuguese.

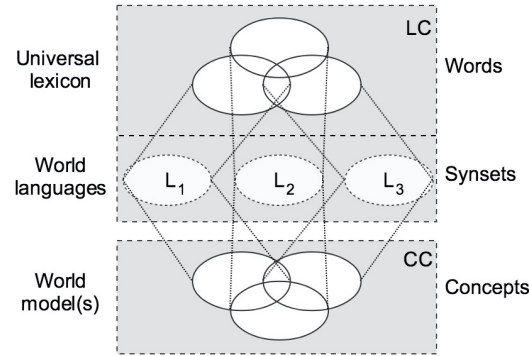
As a result the UKC implements the following stratified theory of meaning:

- 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., [9, 10].

#### 4 World, Languages and Model(s)

The three-layer organization of meaning into words, synsets and concepts, as represented in Figure 2, motivates a three layer design of the UKC, as represented in Figure 3, with the first two layers inside the LC and the third inside the CC. We have:



**Fig. 3.** Languages, Universal lexicon and World model(s).

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

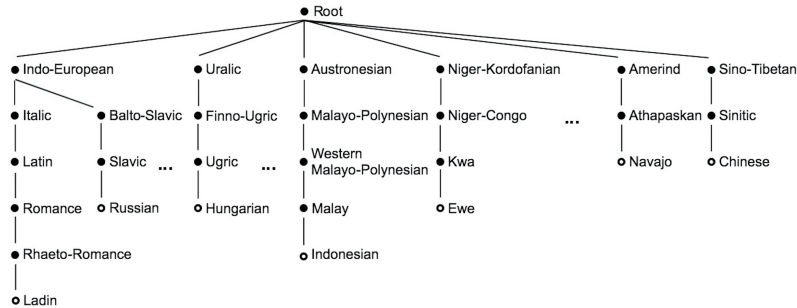
According to a first mainstream interpretation, the CC is the model of the entire world, as it is generated by all the people (speaking all the languages) in the world. However, according to a second interpretation, the CC can also be seen as the union of the models of the world as they are generated by the different people (speaking the different languages) in the world, as represented in Figure 3, e.g., the models of the 7,097 languages registered by the *Ethnologue project*.<sup>4</sup> 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 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.

---

<sup>4</sup> <http://www.ethnologue.com/>

## 5 Language Diversity

The *diversity* of languages appears at many different levels, e.g., phonology, morphology, syntax, and has been the object of extensive studies in the field of *Historical Linguistics* [14] as well as the related field of *Linguistic Typology* [15]. Diversity has many causes, for instance, genetic ancestry (languages which derive from a common language will tend to share more language elements), geographic closeness (people living closer will tend to use many similar or identical words), similarity in culture (people with similar cultures will tend to model the world with similar languages), and so on.



**Fig. 4.** The Phylogenetic Tree of Languages.

We associate the notion of diversity to a set of languages  $\mathcal{L}$ , where the intuition is that the more different the languages, the higher the diversity measure. Thus for instance the diversity measure of {Italian, French} will be smaller than that of {Italian, French, Spanish} and of {Italian, Mongolian}. We take the diversity of the empty set and of the singleton set to be zero. Furthermore, we talk of *relative diversity* (between two languages) when the cardinality of  $\mathcal{L}$  is two. Finally, we talk of *language similarity* (and *relative similarity*) intuitively meaning the opposite phenomenon.

In this paper we concentrate on genetic diversity and adopt the notion introduced in [6] which, in turn, is an evolution of the measure defined in [16]. This notion is based on an analysis of the *Phylogenetic Tree* which describes how languages have progressively descended from other languages [17, 16]. A fragment of this tree is depicted in Figure 4 while Table 2 reports the UKC distribution of languages and continents over phyla (first nine columns), where around 60% of the languages belong to the first four phyla. In Figure 4, the root is just a placeholder for the set of all languages, the intermediate nodes are *language families* or *phyla*, each associated with a set of languages, while the terminal nodes are the actual languages. There are nine children of the root which progressively split to consider all languages. We compute diversity based on the following intuitions:



1. The diversity measure of a language is its distance from the root node. For instance, as from Figure 4, Russian is at distance 4 from the root while Latin is at distance 6.
2. Higher nodes correspond to languages which split before and have evolved independently for more time, thus becoming more diversified. As from [6], this is modeled by associating to each node a weight  $\lambda^{-d}$  with  $\lambda > 1$  and  $d$  being the distance from the root.
3. The diversity of a set of languages, when the set contains at least two languages, is the sum of the diversity of its languages.

The resulting diversity is not normalized. We normalize it by considering the diversity measure of a reference set, e.g., the languages in the Ethnologue project or, as we have done so far, with the diversity measure of the set  $\mathcal{L}_{UKC}$  of the languages of the UKC. We call the resulting two measures *Absolute Diversity* and *(Normalized) Diversity* and we write them as AbsGenDiv and GenDiv, respectively.

Let us consider some examples, computed assuming  $\lambda = 2$ . The Absolute Diversity and Diversity of the languages of the UKC are  $\text{AbsGenDiv}(\mathcal{L}_{UKC}) = 88.127$  and  $\text{GenDiv}(\mathcal{L}_{UKC}) = 1$ , respectively. Furthermore we have, as an example,  $\text{AbsGenDiv}(\{\text{Hungarian, Italian, Polish, Russia, Basque}\}) = 3.469$  and  $\text{GenDiv}(\{\text{Hungarian, Italian, Polish, Russia, Basque}\}) = 0.039$  (with respect to the  $\mathcal{L}_{UKC}$ ).

**Table 2.** Language distributions across phyla.

Phylum	Dep.	Lan.	EU	AS	AM	AF	PA	Example	LanInc( <i>l</i> )	LanQua( <i>l</i> )	#Words
Indo-European	7	115	86	26	1	1	1	English	[0.00; 1.00]	[0.00; 1.00]	[5; 147, 263]
Austronesian	6	36	1	23	2	0	10	Malay	[0.71; 1.00]	[0.16; 0.57]	[8; 24, 081]
Altaic	6	30	16	14	0	0	0	Mongolia	[0.53; 1.00]	[0.25; 0.62]	[12; 86, 574]
Uralic	6	22	22	00	0	0	0	Finnish	[0.01; 1.00]	[0.18; 0.57]	[8; 115, 259]
Niger-Kordofa.	5	21	0	0	0	21	0	Ewe	[0.99; 1.00]	[0.24; 0.60]	[11; 354]
Amerind	4	18	0	0	18	0	0	Navajo	[0.98; 1.00]	[0.18; 0.59]	[10; 1, 460]
Sino-Tibetan	4	18	0	18	0	0	0	Mandarin	[0.10; 1.00]	[0.12; 0.65]	[3; 91, 898]
Afroasiatic	4	14	1	3	0	10	0	Hebrew	[0.91; 1.00]	[0.23; 0.61]	[15; 13, 601]
Caucasian	3	12	9	3	0	0	0	Chechen	[0.97; 1.00]	[0.22; 0.57]	[10; 2, 828]
Creole	3	9	0	0	5	1	3	Tok Pisin	[0.99; 1.00]	[0.22; 0.56]	[9; 485]
small 22 families	4	40	4	11	17	4	4	Basque	[0.94; 1.00]	[0.26; 0.60]	[12; 25, 676]
Total	7	335	139	98	43	37	18	-	[0.00; 1.00]	[0.00; 1.00]	[0; 147, 263]

**Dep.** represents a depth of its corresponding phylum.

**Lan.** 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.

## 6 Resource Quality

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,

**Table 3.** Ten sample languages from Table 2.

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

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 [18], in lexicostatistics [19, 20], mass comparison [21], or the recent work on lexical semantics described in [22]. 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 [6], 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. For lack of space, we provide below a measure of incompleteness and one of incorrectness and exploit them to characterize some aspects of the current state of the UKC. A more complete list of measures will be provided in a follow up longer paper.

### 6.1 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.

$$\text{AbsLanCov}(l) = |\text{Concepts}(l)| \quad (1)$$

$$\text{LanCov}(l) = \frac{|\text{AbsLanCov}(l)|}{|\text{Concepts}(\text{UKC})| - |\text{Gaps}(l)|} \quad (2)$$

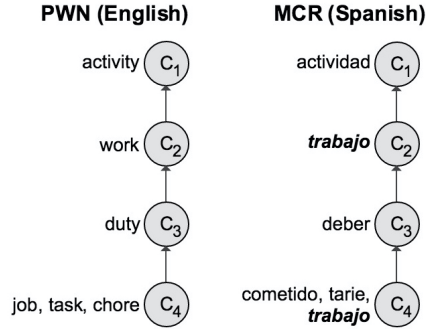
$$\text{LanInc}(l) = 1 - \text{LanCov}(l) \quad (3)$$

where  $\text{Concepts}(l)$  is the set of concepts lexicalized by a language  $l$ ,  $\text{Concepts}(\text{UKC})$  are the concepts in the UKC and  $\text{Gaps}(l)$  are the lexical gaps of  $l$ .  $\text{AbsLanCov}$

is the *Absolute Language Coverage*. Table 2 (column 10), reports the range of values for LangInc in the various phyla, while Table 3 provides its values for ten selected languages. It is interesting to notice how  $\text{LangInc}(\text{English}) = 0.0$ . 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  $\text{LangInc}(\text{Navajo}) = 0.98$ .

## 6.2 Incorrectness

The quality of a language can be measured by several factors, e.g., translation mistakes, wrong senses, and much more. In the following, we analyze the problem of the *psycholinguistic mistakes* which we define as failures of adhering to the principle which, as from [23], 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 provides an example of psycholinguistic mistake in the Spanish WordNet.



**Fig. 5.** A psycholinguistic mistake in Spanish.

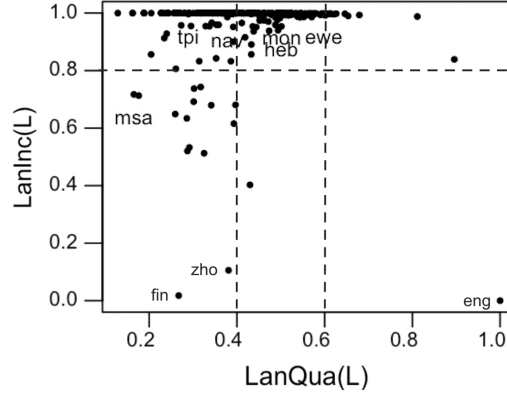
We have the following definitions:

$$\text{AbsLanQua}(l) = -\log_{10}\left(\frac{|\text{PsyMis}(l)| + 1}{|\text{Concepts}(l)|}\right) \quad (4)$$

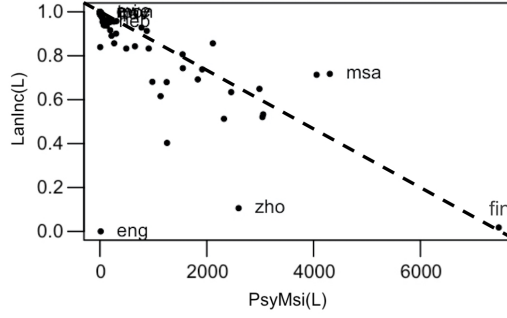
$$\text{LanQua}(l) = \frac{\text{AbsLanQua}(l)}{\text{AbsLanQua}(\text{English})} \quad (5)$$

$$\text{AvgDis}(l) = \frac{\sum_{x \in \text{PsyMis}(l)} \text{dis}(x)}{|\text{PsyMis}(l)|} \quad (6)$$

where  $\text{PsyMis}(l)$  is the set of psycholinguistic mistakes in  $l$ ,  $\text{AbsLanQua}(l)$  and  $\text{LanQua}(l)$  are the *Absolute Language quality* and the *Language quality* of  $l$ , respectively. The number of mistakes varies a lot, going from the fourteen mistakes



**Fig. 6.** Language Incompleteness vs Language Quality.



**Fig. 7.** Language Incompleteness vs Psycholinguistic Mistakes.

of the PWN English to the thousands of mistakes of other languages. 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.  $dis(x)$  is the number of intermediate nodes between two concepts generating the psycholinguistic mistake  $x$ , for instance, in figure 5,  $dis(trabajo) = 2$ . 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.

Figures 6 and 7 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 6 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, the mistakes grow with the size of the language itself. Figure 7 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.

## 7 Related Work

As far as we know, the stratification of meaning into words, synsets and concepts and the resulting three-layer architecture of the UKC has never been proposed before. Relevant work has been done, however, in the development of large scale multilingual resources.

BabelNet [24] is the largest multilingual lexico-semantic resource, obtained from an automatic integration of several resources. Currently, this resource covers 271 languages, 6 million concepts and millions of words. The design decisions underlying BabelNet and the UKC are fundamentally different. The most important difference is that in the UKC we do not allow the addition of entities with the exception of those (a very small minority) which are mentioned in glosses. The main motivation for this decision is that we want to keep the UKC inherently linguistic and focused on concepts. Notice how the number of entities is essentially unbound, order or magnitude bigger than that of concepts and, modulo a (still large) number of exceptions (e.g., the names of famous locations or people), inherently bound to the local cultures. The second main difference is our focus on quantifiable high quality, a property which is hard to maintain when performing automatic resource integration.

Lately, a lot of work has focused on the creation of a *Global Wordnet Grid*, which is currently being instantiated in the *Open Multilingual Wordnet* [25], and whose goal is to link the concepts from different Wordnets. Towards this goal, the Collaborative InterLingua Index (CILI) [26] aims at enabling the coordination among multiple loosely coordinated Wordnets. Finally, as part of the development of Global Wordnet Grid, the work described in [27] has recently introduced the idea of a central registry of concepts. This latter idea is somewhat related to our idea of clustering the synsets with the same meaning under the same concept. However, differently from us, in this work, the different languages are only loosely coupled and there seems no easy way to use the different languages inside a single application. Furthermore, this work seems still somewhat early stage, in terms of number of languages which has been integrated so far, and also, in terms of quality control of the resource.

## 8 The Way Ahead

The UKC is at a mature stage of development, but with still a lot of work to be done. We foresee the following areas of further research: (i) the development of a computational diversity-aware theory of meaning based on the notion of concept defined above, (ii) the use of the UKC for the development of quantitative studies of language diversity, starting from an in depth analysis of lexical gaps, (iii) the further development of the UKC, as a community effort, both in terms of new

languages and of enrichment of the existing languages, (iv) a refinement of the CC aimed to aligning lexical concepts with the results of perception and, last but not least, (v) the extensive use of the UKC in language aware applications, with a focus on large scale video classification.

## Acknowledgments

The first version of the UKC was developed by Ilya Zaihrayeu, around 2004. This implementation was revised many times, most often as a joint effort between Ilya and Marco Marasca. Since the beginning, the UKC has been designed with the goal of supporting the automation of reasoning based on information extracted from text, the original goal being the matching of ontologies [28]. We thank the many postdocs and PhD students who have extensively used the UKC in their research.

The current work is supported by QROWD (<http://qrowd-project.eu>), a Horizon 2020 project, under Grant Agreement No. 732194. The second author is supported by the ESSENCE Marie Curie Initial Training Network, funded by the European Commission’s 7th Framework Programme under grant agreement no. 607062.

## References

1. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. *International journal of lexicography* **3**(4) (1990) 235–244
2. Vossen, P.: Introduction to eurowordnet. *Computers and the Humanities* **32**(2-3) (1998) 73–89
3. Pianta, E., Bentivogli, L., Girardi, C.: Multi-wordnet: developing an aligned multilingual database.”. In: *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January. (2002) 21–25
4. Gonzalez-Agirre, A., Laparra, E., Rigau, G.: Multilingual central repository version 3.0. In: *LREC*. (2012) 2525–2529
5. Von Fintel, K., Matthewson, L.: Universals in semantics. *The linguistic review* **25**(1-2) (2008) 139–201
6. Giunchiglia, F., Batsuren, K., Bella, G.: Understanding and exploiting language diversity. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. (2017) 4009–4017
7. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? *Behavioral and Brain Sciences* **33**(2-3) (June 2010) 61–83
8. Bella, G., Giunchiglia, F., McNeill, F.: Language and domain aware lightweight ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web* (2017)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: *CVPR09*. (2009)
10. Deng, J., Russakovsky, O., Krause, J., Bernstein, M., Berg, A.C., Fei-Fei, L.: Scalable multi-label annotation. In: *ACM Conference on Human Factors in Computing Systems (CHI)*. (2014)

11. Giunchiglia, F., Fumagalli, M.: Concepts as (recognition) abilities. In: Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016). Volume 283., IOS Press (2016) 153
12. Giunchiglia, F., Fumagalli, M.: Teleologies: objects, actions and functions. In: Proceedings of the 36th International Conference on Conceptual Modeling (ER 2017). (2017)
13. Millikan, R.G.: On clear and confused ideas: An essay about substance concepts. Cambridge University Press (2000)
14. Crowley, T., Bower, C.: An introduction to historical linguistics. 4 edn. Oxford University Press (2010)
15. Croft, W.: Typology and universals. Cambridge University Press (2002)
16. Rijkhoff, J., Bakker, D., Hengeveld, K., Kahrel, P.: A method of language sampling. *Studies in Language. International Journal sponsored by the Foundation Foundations of Language* **17**(1) (1993) 169–203
17. Bell, A.: Language samples. universals of human language, ed. by Joseph Greenberg et al., 1.153–202 (1978)
18. McMahon, A., McMahon, R.: Language classification by numbers. Oxford University Press on Demand (2005)
19. Swadesh, M.: Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics* **21**(2) (1955) 121–137
20. Swadesh, M.: The origin and diversification of language. Transaction Publishers (1971)
21. Greenberg, J.H.: Universals of language. (1966)
22. Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J.F., Maddieson, I., Croft, W., Bhattacharya, T.: On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* **113**(7) (2016) 1766–1771
23. Miller, G.A.: Nouns in wordnet: a lexical inheritance system. *International journal of Lexicography* **3**(4) (1990) 245–264
24. Navigli, R., Ponzetto, S.P.: Babelnet: Building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics (2010) 216–225
25. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: *ACL* (1). (2013) 1352–1362
26. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: Cili: the collaborative interlingual index. In: Proceedings of the Global WordNet Conference. Volume 2016. (2016)
27. Vossen, P., Bond, F., McCrae, J.: Toward a truly multilingual globalwordnet grid. In: Proceedings of the Eighth Global WordNet Conference. (2016) 25–29
28. Giunchiglia, F., Autayeu, A., Pane, J.: S-match: an open source framework for matching lightweight ontologies. *Semantic Web* **3**(3) (2012) 307–317