ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 9, Number 2
december 2023

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000945

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_9_2

IJCoL

Volume 9, Number 2
december 2023

CONTENTS

#DEACTIVHATE: An Educational Experience for Recognizing and Counteracting Online Hate Speech
Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Simona Frenda, Viviana Patti 7

Towards Cross-lingual Representation of Prototypical Lexical Knowledge
Francesca Grasso, Luigi Di Caro 33

The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German
Aivars Glaznieks, Jennifer-Carmen Frey, Andrea Abel, Lionel Nicolas, Chiara Vettori 53

Intelligent Natural Language Processing for Epidemic Intelligence
Danilo Croce, Federico Borazio, Giorgio Gambosi, Roberto Basili, Daniele Margiotta, Antonio Scaiella, Martina Del Manso, Daniele Petrone, Andrea Cannone, Alberto Mateo Urdiales, Chiara Sacco, Patrizio Pezzotti, Flavia Riccardo, Daniele Mipatrini, Federica Ferraro, Sobha Pilati 77

POS Tagging and Lemmatization of Historical Varieties of Languages. The Challenge of Old Italian
Manuel Favaro, Marco Biffi, Simonetta Montemagni 99

Towards Cross-lingual Representation of Prototypical Lexical Knowledge

Francesca Grasso*
Università degli Studi di Torino

Luigi Di Caro**
Università degli Studi di Torino

In order to be concretely effective, many Natural Language Processing (NLP) applications require the availability of lexical resources providing varied, broadly shared, and language-unbounded (i.e., not limited to any specific language or linguistic system) lexical information. However, state-of-the-art knowledge models typically focus on specific levels of semantic analysis rather than adopting such a comprehensive and cross-lingual approach to lexical knowledge. This is often due to the theoretical paradigms on which such resources are based, each addressing the semantic phenomenon from a (de)finite perspective. In this paper, we first suggest a maximalist approach to lexical semantics to pursue through the idea of semantic prototype and linguistic representativeness as easily applicable to textual corpora. Starting from this conceptual framework, we thus propose a novel corpus-based automatable methodology for knowledge modeling based on a multilingual word alignment mechanism. This model enables the retrieval and encoding of prototypical, language-unbounded, and naturally disambiguated lexical knowledge in the form of diversified conceptual links between words and their senses. Results from a simple implementation of the proposal show relevant outcomes that are not found in other resources. Finally, different application opportunities of the proposed model will be presented.

1. Introduction

The exploitation of lexical resources constitutes a key issue for several Natural Language Processing tasks and applications, such as Word Sense Disambiguation and Machine Translation. However, their potential may vary widely depending on the nature of the lexical-semantic knowledge they encode, as well as on how the linguistic data are stored and linked within the given lexical network (Zock and Biemann 2020). Extra-linguistic (i.e. *encyclopedic*) information in particular has traditionally constituted a neglected area in the field of Knowledge Modeling due to the challenging nature of its encoding. In order to deal with the complexity and fluidity of lexical semantics, computational approaches typically deconstruct the phenomenon into less complex units that are easier to manipulate. That is, the "*dividi et impera*" strategy adopted towards linguistic phenomena serves to create a streamlined and functional structure for narrow detailed analyses of specific language levels or phenomena (Petricca 2019). As a result, lexical resources often fail to return more comprehensive and context-sensitive lexical-semantic information.

The resources that are presently available, such as WordNet (Miller 1995), typically encode language-bounded lexical-semantic knowledge mainly in terms of word senses, defined by textual (i.e. *dictionary*) definitions, and lexical entries are linked and put in

* Dept. of Computer Science - Corso Svizzera 185, 10149 Turin, Italy. E-mail: fr.grasso@unito.it
** Dept. of Computer Science - Corso Svizzera 185, 10149 Turin, Italy. E-mail: luigi.dicaro@unito.it

context through lexical-semantic relations. These relations, being primarily of a paradigmatic nature, are characterized by a sharing of the same defining properties between the words and a requirement that the words be of the same syntactic class (Morris and Hirst 2004). Typically related words are therefore not represented due to the absence of syntagmatic links (e.g., co-occurrences) and other untyped relations such as free-associations (Nelson, McEvoy, and Dennis 2000; Nelson, McEvoy, and Schreiber 2004). Additionally, word senses suffer from a lack of explicit common-sense knowledge and context-dependent information. Finally, the well-known fine granularity of word senses in WordNet (Palmer, Dang, and Fellbaum 2007) is due to the lack of a meaning encoding system capable of representing concepts in a flexible way. Other kinds of resources such as FrameNet (Baker, Fillmore, and Lowe 1998) and ConceptNet (Speer, Chin, and Havasi 2017) embrace a more inclusive perspective of the semantic phenomenon, since they return empirically-retrieved lexical material, i.e., real-world language data collected from naturally-occurring sources. However, while providing insight into extra-linguistic information, these models still lack flexibility and deliver different types and degrees of structured (monolingual) semantic information and disambiguation capabilities, failing to capture the multiple layers of knowledge associated to meaning units.

In this contribution, we first discuss an alternative theoretical framework that encompasses a maximalist (i.e. comprehensive) view of the semantic phenomenon, claiming that even a computational (i.e.formalist) approach to semantics cannot overlook a broader concept of knowledge embracing different levels of analysis, as this is crucial for many NLP applications. Both dictionary and encyclopedic views of lexical knowledge will be therefore taken into account, together with a brief overview and problematization of the different theoretical perspectives from which they arise. The ideas of semantic prototype and linguistic representativeness (as applied to textual corpora) will be finally addressed.

Drawing from this conceptual background, we propose an original, corpus-based methodology for the retrieval and representation of prototypical, language-unbounded, naturally disambiguated lexical-semantic information that relies on a multilingual word alignment mechanism. Starting from the conception of textual corpora as a key tool to access empirical and representative -therefore prototypical - lexical knowledge, we leverage this textual property by exploiting corpora in k different languages in order to acquire and align varied lexical-semantic material in the form of k-Multilingual Concept (MC^k) (Grasso, Lovera Rulfi, and Di Caro 2022). MC^ks consist of multilingual alignments of semantically equivalent words in k different languages, that are generated through a defined linguistic context and linked via empirically determined semantic relations without the use of any sense disambiguation process.

As a third contribution, we present an implementation of the methodology that allows the automatic acquisition of MC^ks from several corpora in three languages (English, Italian, and German). To evaluate the effectiveness of our methodology, we examined our knowledge acquisition system's ability to uncover new lexical relations that had not been previously identified by a state-of-the-art resource (BabelNet (Navigli and Ponzetto 2010)). It should be noted that our aim is not to overcome any existing resource, but instead to integrate new, unbiased[1] semantic relations from a novel multilingual alignment mechanism. As the results of the implementation will show, this

1 In this paper, when we refer to 'bias', we mean the influence of language-specific elements and lexicographic idiosyncrasies that can be observed in individual resources. The term is defined and discussed in more detail in Section 3.1.

method enables the encoding of varied layers of lexical knowledge, in terms of both syntagmatic and paradigmatic relations, providing networks of diversified conceptual links between words in - and shared by - different languages. This system, therefore, enhances the encoding of prototypical semantic information of concepts that is also likely to be free from strong cultural-linguistic specificities and lexicographic biases.

The benefits provided by our novel multilingual word alignment mechanism are thus fourfold: *(i)* a linguistic and lexicographic de-biasing of lexical knowledge; *(ii)* naturally-disambiguated aligned lexical items; *(iii)* the discovery of novel lexical-semantic relations; and *(iv)* the representation of prototypical semantic information of concepts in- and shared by different languages.

2. Related Work

On one side, lexicons are built on top of synsets[2] and contextualize meanings (or senses) mainly in terms of paradigmatic relations. WordNet (Miller 1995) and BabelNet (Navigli and Ponzetto 2010) can be seen as the cornerstone and the summit in that respect. However, if on the one hand, WordNet's dense network of taxonomic relationships allows a high degree of systematization, on the other hand, a key unsolved issue with "*wordnets*" is the fine granularity of their inventories. Note that multilingualism in BabelNet is provided as an indexing service rather than as an alignment and unbiasing systematization method. While these projects address products of linguistic structure such as paradigmatic and syntagmatic relations, other works focus on extra-linguistic aspects of language, such as associative relations (Nelson, McEvoy, and Schreiber 2004; Buchanan et al. 2013). Extensions of these resources also include Common-Sense Knowledge (CSK), which refers to some (to a certain extent) widely accepted and shared information. CSK describes the kind of general knowledge material that humans use to define, differentiate, and reason about the conceptualizations they have in mind (Ruggeri, Di Caro, and Boella 2019). ConceptNet (Speer, Chin, and Havasi 2017) is one of the largest CSK resources, collecting and automatically integrating data starting from the original MIT Open Mind Common Sense project[3]. However, terms in ConceptNet are not disambiguated. Property norms (McRae et al. 2005; Devereux et al. 2014) represent a similar kind of resource, which is more focused on the cognitive and perception-based aspects of word meaning. Norms, in contrast with ConceptNet, are based on semantic features empirically constructed via questionnaires producing lexical (often ambiguous) labels associated with target concepts, without any systematic methodology of knowledge collection and encoding. Another widespread modeling approach is based on vector space models of lexical knowledge. Vectors are automatically learned from large corpora utilizing a wide range of statistical techniques, all centered on Harris' distributional assumption (Harris 1954), i.e. words that occur in the same contexts tend to have similar meanings. Well-known models include word embeddings (Mikolov et al. 2013; Pennington, Socher, and Manning 2014; Bojanowski et al. 2016), sense embeddings (Huang et al. 2012; Iacobacci, Pilehvar, and Navigli 2015; Kumar et al. 2019), and contextualized embeddings (Scarlini, Pasini, and Navigli 2020). However, the relations holding between vector representations are not typed (i.e., they are not explicitly categorized based on any specific type of information) nor are they systematically organized.

2 Words considered as synonyms in specific contexts.
3 https://www.media.mit.edu/projects/open-mind-common-sense/overview/

Among the several other modeling strategies proposed, lexicographic-centered resources have been focused on the contextualization of lexical items within syntactic structures, e.g. Corpus Pattern Analysis (CPA) (Hanks 2004), situation frames such as FrameNet (Fillmore 1977; Baker, Fillmore, and Lowe 1998) and conceptual frames (Moerdijk, Tiberius, and Niestadt 2008; Leone et al. 2020). Words are not taken in isolation and the meaning they are attributed is connected to prototypical patterns or typed slots, i.e., specific roles or semantic relationships that are associated with words or phrases within a frame. However, these theories and methods for building semantic resources remain linked to the lexical basis and do not manage the mentioned biases.

The problem of identifying the correct meaning of words depending on the context of occurrence represents one of the oldest tasks in the field of Natural Language Processing. The process of Word Sense Disambiguation hides a wide range of complexities, such that even after decades of technological advancement the current state of the art is still far from reaching more-than-good accuracy levels (Lacerra et al. 2020). Many studies have already proved the advantages of a cross-lingual approach to Word Sense Disambiguation (Brown et al. 1991; Apidianaki 2013; Chan and Ng 2005; Diab and Resnik 2003). The use of translations of a given word as sense labels avoids the need for manually created sense-tagged corpora and sense inventories. Moreover, a cross-lingual approach deals with the sense granularity problem: finer sense distinctions became truly relevant as far as they get lexicalized into different translations of the word (Lefever and Hoste 2013). However, existing works usually exploit either parallel texts or multilingual Wordnets, therefore relying on an intrinsically limited number of de-facto already built alignments.

3. Background and Motivation

In this section, we outline the theoretical framework behind our contribution and the motivation that guided its development. First, a brief overview of the biases that typically affect lexical knowledge encoding will be provided. Then, we problematize different theoretical perspectives underlying existing resources. Finally, the idea of semantic prototype and its application to textual corpora will be discussed.

3.1 Bias types

Lexical knowledge provided by lexical resources - especially when monolingual - will inherently carry different types of biases. In particular, *i)* language-bound elements and *ii)* lexicographic biases affect the encoding, consumption, and exploitation of lexical knowledge in downstream tasks.

Language specificity. Lexical information encoded in a language's lexicon, as well as the potential contexts in which a given lexeme can occur, inevitably reflect the socio-cultural background of the speakers of that language. Lexical resources used for the compilation of lexical knowledge are often conceived as monolingual, therefore they mostly return culture-bounded semantic information which does not account for more shared knowledge.

Lexicographic bias. The nuclear components extracted from textual definitions can be different depending on the resource used, even within a single language (Kiefer 1988). For example, the definition of "*cow*" reported by the Oxford Dictionary is "*a large animal kept on farms to produce milk or beef*" while the Merriam-Webster Dictionary reports "*the*

mature female of cattle". Both endogenous and exogenous properties can be subjectively reported (Woods 1975), such as the term "*large*" and the milk production respectively.

3.2 Type of Knowledge (Models)

Methodologies underpinning state-of-the-art knowledge models draw from theoretical backgrounds whose scope of investigation is narrowed according to their corresponding specific view of (lexical) knowledge. As a consequence, they lack an inclusive and broader approach to semantics. Depending on the type of resource, the description of lexical meaning typically involves the encoding of either just linguistic (dictionary knowledge) or extra-linguistic material (encyclopedic knowledge), although rarely both. This minimalist plan of action fails to return a comprehensive and language-independent description of a given concept, even when it manages polysemy. Accordingly, current knowledge models can be seen as divided into two macro-categories, based on the type of knowledge they intend to encode:

i. Resources displaying dictionary (linguistic) knowledge, mirroring a formalist (mainly structuralist and generative) approach to semantics. In the dictionary view of meaning, there is a separation of core meaning (semantics) from non-core actual meaning (pragmatics) (Kecskes 2012). As a consequence, the description of meaning can only be of a purely linguistic nature (Kiefer 1988). A well-known model mirroring this kind of approach to lexical semantics is, for instance, WordNet (Miller 1995). The fine granularity of this kind of resource and the absence of encyclopedic knowledge, while allowing a high systematization of the linguistic data, determines an artificial simplification that does not always reflect empirical reality (meant as the actual, real-world meaning and usage of words and language).

ii. Models embracing a cognitive (i.e. maximalist) approach to semantics and promoting the encoding of encyclopedic knowledge. According to the cognitivist perspective, the definition of lexical meaning requires a reference to our mental representations of concepts and to the encyclopedic knowledge they embody (Fauconnier 1997; Tulving 1983). Encyclopedic meaning arises in context(s) of use: the "selection" of actual situational meaning is informed and maybe even determined by contextual factors (Kecskes 2012; Evans 2006). According to this approach, there is no definable, pre-existing word meaning because the meaning of a word in context is selected and shaped by encyclopedic knowledge. Models adopting this view of semantics often lack systematization and do not return clear and sorted lexical information, but rather generic and approximate hints regarding the given concept. For example, FrameNet (Fillmore 1977; Baker, Fillmore, and Lowe 1998) lacks a concrete reference to the encyclopedic element. This is due to its connection with the "strong" version of the cognitivist approach which no longer distinguishes linguistic material from extra-linguistic one (Petricca 2019).

While avoiding radical adherence to any preselected theoretical framework, it becomes clear that the observation of semantic phenomenology requires the consideration of other than the mere textual definition of a word or its idiosyncratic collocation in a context. Therefore, we claim that the representation of both dictionary and encyclopedic knowledge (a concept that in the NLP field overlaps to a certain extent with *commonsense* knowledge) should be of paramount importance for a comprehensive linguistic resource.

3.3 Semantic prototype and Representativeness

Due to its complex and fluid nature, lexical semantics necessarily needs to undergo a process of abstraction and simplification to be encoded into a formal model. The selection of a limited yet representative and detectable semantic material leads to taking into account the idea of *semantic prototype*. This concept is borrowed from psychology and cognitive semantics, approaches that have traditionally looked at prototypicality and salience as key concepts in the study of semantics (Lakoff 1987; Rosch 1975). Prototypical information is the semantic material that is perceived as significant and representative by the speakers and is part of a shared knowledge (Hampton 2015). Therefore, it cannot correspond to the mainstream textual definition of the given (lexical) unit, nor can it be left to the free interpretation of the speaker. The notion of semantic prototype, in line with what has been stated in the previous section, recalls once again the inclusion of both linguistic- and extra-linguistic knowledge. In this regard, textual corpora represent a key yet simple tool for investigating the dimensions of lexical semantics that lie behind the mere dictionary knowledge of words. As is known, the use of corpora provides a solid empirical foundation for general-purpose language tools and descriptions and enables analyses of a scope not otherwise possible (Biber 1993). The concept of prototype as conveying the idea of *representativeness* can be easily applied to textual corpora since they are typically built to carry out research on a set of texts that is as representative as possible of a target population of interest (Egbert, Larsson, and Biber 2020; Hunston 2002). The point of a corpus is precisely to be an accurate representation of that target register, dialect, or entire language (McEnery, Xiao, and Tono 2006). This leads to consider that the use of many textual corpora in multiple languages may widen the concept of representativeness to a degree where it is possible to capture and eventually retrieve language-independent prototypical lexical knowledge, meant as semantic information of concepts in- and shared by different languages. Besides possibly overcoming the issue of *language-specificity* through leveraging a set of differently-built language corpora, by doing so we can also minimize *lexicographic biases*. The methodology presented in this work builds upon the above-outlined assumptions.

4. The multilingual word alignment

As is known, a single word form can be associated with more than one related sense, causing what is referred to as semantic ambiguity, or polysemy. This phenomenon, however, manifests itself differently across languages, since each language encodes meaning into words in its own particular way. Therefore, it may decrease when putting lexical items in a reciprocal relation, i.e., when aligned. While a given language may provide only a single disambiguation context for a word, the use of parallel languages may indeed help further restrict word sense variability (Atkins, Fillmore, and Johnson 2003). For example, the concept of "*discharge from an office or position*" may be encoded into the English verb form "*to fire*" which is however highly ambiguous, counting twelve different verbal senses in WordNet. The same concept is expressed by another polysemous term in Italian, i.e. "*licenziare*". However, the words *fire - licenziare* when associated with each other represent a bilingual encoding of that single concept which naturally avoids ambiguity, given that there are no other meanings that the two words may share. Thus, translations of a target word into one or more languages provide it a disambiguation context and may serve as sense labels (Lefever and Hoste 2013).

Based on this assumption, it is possible to exploit this cross-language property to disambiguate a given word using its semantic equivalent in another language when they both occur in the same context.

Accordingly, we developed a corpus-based knowledge acquisition methodology that features the power of word sense disambiguation, relying on a multilingual alignment mechanism. Many works (Brown et al. 1991; Chan and Ng 2005; Apidianaki 2013; Lefever and Hoste 2013; Diab and Resnik 2003), have already shown the advantages of multilingual word alignments to perform Word Sense Disambiguation, although dwelling on the exploitation of either parallel corpora or multilingual wordnets, i.e, on already existing and pre-determined cross-lingual lexical material. In this work, we propose to leverage this property of languages through a more dynamic system featuring a broader scope.

After providing a brief illustration of the languages we involved in this first phase of the project, we describe more in detail the methodology by using a basic example. Afterwards, an implementation of the proposed mechanism is presented.

4.1 Languages involved

Among the benefits provided by the multilingual word alignment methodology we propose, one is that it prevents the represented lexical information from containing strong language-dependent information. This objective is pursued through the use of three different languages, reflecting in turn three diverse backgrounds. For this first trial, we involved English, German, and Italian. These languages were chosen primarily because we are proficient in them, therefore we are able to exert control over the data, as well as to interpret the results properly. Concurrently, given the nature of the methodology, it was necessary to select a set of languages with a certain degree of similarity in terms of shared lexical-semantic material. Indeed, the alignment mechanism can work and be effective as long as the lexical-semantic systems of the languages involved reflect a somehow similar cultural-linguistic background. For example, we might expect languages to agree on the meanings of "carp", "cottage" and "sled" as long as speakers of these languages have comparable exposure to the relevant data. We would not expect a language spoken in a place without carps to have a word corresponding to "carp". The purpose of this project is not to forcibly identify universally valid semantic relationships, but rather to not report biased information deriving from the use of data coming from a single linguistic context. For this reason, in our case, the choice fell on European languages [4] (two Germanic languages and a Romance one).

4.2 Method

We now describe in detail the alignment mechanism through a basic example. Consider the following word forms: *wool* (EN); *Wolle* (DE); *lana* (IT), expressing a single target concept[5].

For each of the three lexical forms we collect[6] a set of related words in terms of paradigmatic (e.g. synonyms) and syntagmatic (e.g. co-occurrences) relations by inspecting differently-built textual corpora. The target-related words can possibly be

4 By "European" we refer to the European linguistic area.
5 An absolute monosemy is, of course, realistically unreachable.
6 The detailed process of the data gathering procedure is explained in section 5.

Table 1
Unordered lists of single-language related words for <*wool* (EN), *Wolle* (DE), *lana* (IT)>.

wool	**Wolle**	**lana**
sheep	*Schal*	*cotone*
cotton	*spinnen*	*Biella*
synthetic	*Baumwolle*	*sintetica*
spin	*Rudolf*	*sciarpa*
scarf	*synthetisch*	*pecora*
mitten	*Schafe*	*filare*

Table 2
Examples of aligned concept-related words for <*wool* (EN), *Wolle* (DE), *lana* (IT)>.

wool		**Wolle**		**lana**
sheep	$\leftrightarrow$	*Schafe*	$\leftrightarrow$	*pecora*
cotton	$\leftrightarrow$	*Baumwolle*	$\leftrightarrow$	*cotone*
synthetic	$\leftrightarrow$	*syntetisch*	$\leftrightarrow$	*sintetica*
spin	$\leftrightarrow$	*spinnen*	$\leftrightarrow$	*filare*
scarf	$\leftrightarrow$	*Schal*	$\leftrightarrow$	*sciarpa*

modifiers, verbs, or substantives. We thus obtain three different lists of words, one for each of the languages involved. The retrieved terms in the lists are still potentially ambiguous since they refer to a lexical form rather than a contextually defined concept. Table 1 provides a small excerpt of such unordered lists of related words. As can be noted, the lexical data in the lists consist of either co-occurrence, synonyms, or other related words belonging to the same semantic category of the target word. The items in the lists do not provide any other information besides their relation to the target word, and the lists are unrelated to each other.

The lexical data in the lists are subsequently compared and filtered by means of a two-way translation step in order to select only the semantic items that occur in all the lists, i.e., those shared by the three languages, in the reported example. The resulting words are thus aligned with their semantic counterparts, generating a set of semantically equivalent aligned triplets, as shown in Table 2.

As can be seen, the items in the three lists are now re-ordered and aligned according to semantic criteria. That is, they are represented as related to each other through a semantic equivalence relation. $\leftrightarrow$ is used to symbolize this semantic correlation.

This multilingual word alignment provides, as a consequence, an automatic Word Sense Disambiguation system. Once the triplets are formed, their members will be indeed associated with a likely unique sense, i.e. the one coming from the intersection of all possible language-specific senses related to the three words. In other terms, the target-related words, once aligned, naturally identify (and provide) a common semantic context. As a consequence, potentially polysemous words are disambiguated through such context, without any support from sense repositories.

For example, the context-consistent sense of the verb *to spin* (EN), which is a highly polysemous word in English, can be identified by selecting the only sense that is also

shared by the other two aligned words, i.e. "*turn fibres into thread*". In fact, neither *spinnen* (DE) nor *filare* (IT) can possibly mean e.g. "rotate".

The three words are thus disambiguated due to the mutual support and interaction from the reciprocal language systems, as depicted by the representation in Figure 1. The diagram shows the process of a bilingual alignment between a polysemous word W_x in a language $L1$ and two different word forms in a further language $L2$. By pairing the two meanings of W_x, referred to as $sense - A$ and $sense - B$, with their corresponding lexicalizations in $L2$ (W_y and W_z), the polysemous word in $L1$ can be disambiguated through the support of $L1$ words. Through this mechanism, it is possible to create multilingual word connections that are able to disambiguate, enrich, and possibly reassemble senses in the referenced repositories.

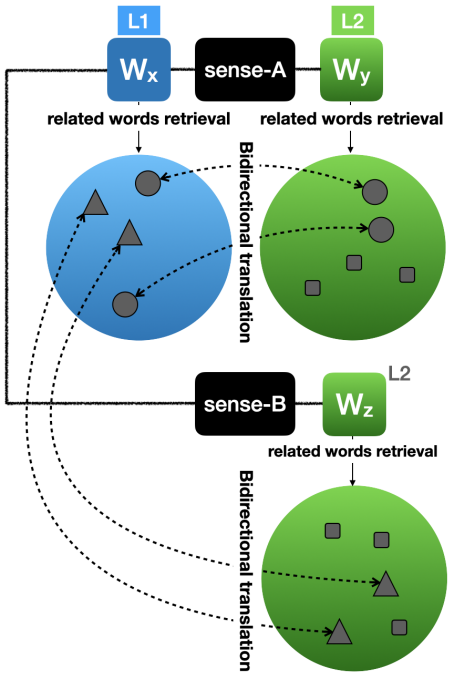

Figure 1
Simple sketch of bilingual alignment for a polysemous word W_x in $L1$-language. Through a different lexicalization of the two meanings A and B into a further $L2$-language (W_y and W_z in $L2$), it is possible to create multilingual word connections able to disambiguate, enrich, and possibly reassemble senses in the referenced repositories.

This mechanism generates a twofold effect: besides performing word sense disambiguation, it also provides lexical knowledge in the form of (paradigmatic and syntagmatic) lexical-semantic relations between words that is also language-unbounded. In the first place, the uncontrolled character of the data retrieval and alignment process offers the generation of novel lexical-semantic relations that are likely not available in other structured resources. Additionally, since the resulting set of words related to the target can be only the one shared by multiple languages, the lexical knowledge it encodes does not reflect a single cultural/linguistic background, but rather a common and shared one. For example, in Table 1 the presence of the word "*Biella*" among the list of words related to "*lana*", probably refers to the fact that the Italian city Biella is (locally) famous for its wool, therefore the two words may co-occur frequently. Similarly,

if we consider the alignment <*cat* (EN), *Katze* (DE), *gatto* (IT)>, a lexeme related to the English word form would be "*rain*", due to the well-known idiom "*it's raining cats and dogs*". However, neither "*Biella*" nor corresponding words for "*rain*" can possibly result in the lists of related words of the respective other languages, being language-specific items within those contexts. Therefore, the lexical information provided by the alignment mechanism will be free from strong language-bounded information. Finally, as illustrated in the next section, by exploiting multiple and differently built resources, we are able to reduce arbitrariness and lexicographic biases within the lexical knowledge represented.

4.3 k-Multilingual Concepts

As suggested in (Grasso, Lovera Rulfi, and Di Caro 2022), we can easily refer to the multilingual word alignments as instances of *k-Multilingual Concept* (hereinafter MC^k), which consists of a concatenation of k lexical items referring to a single concept in k different languages. MC^k constitutes a novel lexical-semantic encoding model bridging between words and senses that is based on the above-described cross-lingual alignment in k different languages. For example, if we consider the concept "*cat*" (as "*domestic cat*"), its $MC^{EN,IT,DE}$ for the three languages English, Italian, and German would be:

$$cat^{EN} \oplus gatto^{IT} \oplus Katze^{DE}$$

where the symbol $\oplus$ represents a simple concatenation operator. As can be noted, MC^ks are basically pseudowords that result from (and consist of) the alignment of multilingual, semantically equivalent lexical forms of a given concept.

5. Implementation

In this section, we describe the details and results of an implementation of the proposed alignment mechanism. It consists of the automatic acquisition of prototypical disambiguated and unbiased lexical information from language-specific corpora in the form of k-Multilingual Concepts. In particular, the system is composed of two main modules: context generation and an alignment procedure. We finally report the results of an evaluation to highlight mainly *(i)* the autonomous disambiguation power of the approach, *(ii)* the quality of the alignments and their unbiased, shared and syntagmatic nature, and *(iii)* the amount of unveiled lexical-semantic relations not covered by existing state-of-the-art resources such as BabelNet.

5.1 Context for multilingual alignment

To start an automatic MC^k extraction process for a given concept C the first requirement is to have a seed, i.e., a MC^k *head* that is constituted by k word forms representing C, one for each language. Once the MC^k *head* has been formed, we use Sketch Engine (Kilgarriff et al. 2014), a corpus management engine, to obtain lists of words related to each single word form that makes up the MC^k *head*, as shown in the example in Table 1. We employ three families of non-semantically annotated large corpora to search for related words in the three languages: the TenTen corpora containing 10+ billion words of generic web content (Jakubíček et al. 2013), the TJSI corpora composed of news articles

Table 3
10 automatic alignments (out of 74) for the target concept <*scale* (EN), *bilancia* (IT), *Waage* (DE)> (BabelNet synset:00069470n).

POS	scale	bilancia	Waage
noun	accuracy	precisione	Genauigkeit
noun	balance	equilibrio	Balance
noun	bulk	massa	Masse
noun	control	controllo	Kontrolle
noun	device	dispositivo	Gerät
noun	figure	cifra	Zahl
adj	accurate	preciso	genau
adj	smart	intelligente	intelligent
verb	indicate	indicare	zeigen
verb	set	regolare	einstellen

(Trampuš and Novak 2012)[7] and the EUR-Lex legal corpora (Baisa et al. 2016). Then, we merge the retrieved related words in the three target languages obtaining three lists (hereinafter *EN*-list, *IT*-list, and *DE*-list), each divided into four categories: *i)* similar nouns, *ii)* co-occurring nouns, *iii)* co-occurring adjectives and *iv)* co-occurring verbs. Finally, we assign a weight to each related word by directly importing the built-in scores of Sketch Engine tools, which are based on the logDice coefficient, as detailed in (Kilgarriff et al. 2014).

5.2 Multilingual alignment

To obtain the MC^ks alignments like those shown in Table 2 we search for cross-match translations using the PanLex API[8], which is focused on words rather than on sentences, and the Google Translate API[9]. In particular, we take each related word t^{EN}, category by category, from the *EN*-list and query the API to get their possible translations into the other two languages (IT, DE). We then try to match each translated item with the previously-retrieved sets of related words in *IT*, *DE*-lists. Whenever the [$t^{EN} \leftrightarrow t^{IT}$]; [$t^{EN} \leftrightarrow t^{DE}$] match succeeded, we finally check any possible [$t^{IT} \leftrightarrow t^{DE}$] match. If a [$t^{EN} \leftrightarrow t^{IT} \leftrightarrow t^{DE}$] semantic equivalence occurs, then the alignment can take place and it will constitute a $MC^{EN,IT,DE}$. Table 3 shows a selection of automatic alignments for the concept *scale* (bn:*00069470n*). Finally, we assign a score to each $MC^{EN,IT,DE}$ by averaging the SketchEngine scores of the three related words.

As last step, we associate BabelNet synsets (always those directly linked to WordNet synsets, if present) and WordNet synsets to the alignments. Specifically, we find the n synsets that have all the given three word forms in the three languages. One of the following three cases may hence occur:

7 TJSI stands for Timestamped JSI web corpus; JSI, in turn, refers to the Jozef Stefan Institute, the institution that provided the corpora. TJSI versions used: English (60+ billion words), Italian (8.4+ billion words), German (6.9+ billion words).
8 https://dev.panlex.org/api/.
9 `https://cloud.google.com/translate`.

- $n = 1$, meaning that the $MC^{EN,IT,DE}$ corresponds to a completely disambiguated concept;
- $n > 1$, when multiple synsets may be associated with a single [$t^{EN} \leftrightarrow t^{IT} \leftrightarrow t^{DE}$] triplet;
- $n = 0$, in case no existing BabelNet synset or WordNet synset actually connects the three word forms.

It is interesting to note that the last two cases cover different situations, such as a missing synset encoding a specific concept ($n = 0$, e.g. significant for sense induction) or overlapping synsets ($n > 1$, e.g. useful for sense clustering). Table 4 shows an excerpt of automatically-generated knowledge around the $MC^{EN,IT,DE}_{book-written\ work}$ *head*: the table also reports the relatedness score and the available BabelNet synset(s) (if present) associated to each triplet. The label *rel* as heading of the corresponding column stands for *related words*, while *c* and *s* indicate the type of related word, being abbreviations for *co-occurrence* and *similar (nouns)*, respectively. As can be noted, the column *synset(s)* of the table depicts the three different scenarios previously described: the occurrence of a single synset for a given $MC^{EN,IT,DE}$ (as e.g. for $read \oplus leggere \oplus lesen$) indicates that the concept expressed by the triplet is fully disambiguated in BabelNet; when a triplet is linked to more than one synset, as with $title \oplus titolo \oplus Titel$, this implies that the $MC^{EN,IT,DE}$ maintains a certain degree of ambiguity; finally, the $MC^{EN,IT,DE}$s with no available synset, as in the case of $thought \oplus pensiero \oplus Gedanke$, constitute new elements not yet covered in the BabelNet semantic network.

5.3 Evaluation

In this section, we briefly present a task designed to demonstrate the validity of the proposed approach. We elucidate the evaluation method using a step-by-step procedure. Our work was motivated by the desire to improve the accuracy and completeness of semantic relationships between words in different languages. We, therefore, do not aim to overcome state-of-the-art resources such as BabelNet, but rather to incorporate new, unbiased semantic relations from a novel multilingual alignment mechanism.

To assess the effectiveness of our methodology, we aim to understand the extent to which our knowledge acquisition system can unveil lexical relations yet uncovered by a state-of-the-art resource (BabelNet). Our primary goal is to gain an understanding of the related words associated with a given concept in BabelNet, identified by a synset, and compare them with the results produced and aligned by our automated methodology for the same synset. This comparison allows us to examine the degree of overlap or dissimilarity between our results and those of the state-of-the-art resource. The following steps outline the comparison process:

- First, we generate sets of related words from Babelnet for a specific synset.
- Using the BabelNet API, we retrieve English, Italian, and German lexicalizations of the associated BabelNet synset, along with the glosses related to them.

Table 4
Fragment of automatically-generated multilingual alignments ($MC^{EN,IT,DE}$s) for the concept *book* (WordNet synset *book - a written work or composition that has been published*), over the three languages.

$MC^{EN,IT,DE}_{book-written\ work}$								
	EN		IT		DE	score	rel	synset(s)
head	*book*	⊕	*libro*	⊕	*Buch*			
nouns								
↓	reading	⊕	lettura	⊕	Lesen	0.580	c	bn:66372n
	author	⊕	autore	⊕	Autor	0.592	c	bn:7287n
	title	⊕	titolo	⊕	Titel	0.130	s	bn:77409n (...)
	chapter	⊕	capitolo	⊕	Kapitel	0.121	s	bn:182115n
	text	⊕	testo	⊕	Text	0.503	c	bn:76732n (...)
	topic	⊕	argomento	⊕	Thema	0.330	s	bn:74900n
	editor	⊕	editore	⊕	Herausgeber	0.464	c	bn:15417659n
	paper	⊕	carta	⊕	Papier	0.451	c	bn:60464n
	library	⊕	biblioteca	⊕	Bibliothek	0.430	c	bn:50968n
	thought	⊕	pensiero	⊕	Gedanke	0.345	s	*<no synset>*
...	...		...		...	...	...	...
verbs								
↓	read	⊕	leggere	⊕	lesen	0.678	c	bn:92426v
	write	⊕	scrivere	⊕	schreiben	0.606	c	bn:93281v (...)
	sell	⊕	vendere	⊕	verkaufen	0.532	c	bn:93472v
	buy	⊕	acquistare	⊕	kaufen	0.484	c	bn:84331v
	illustrate	⊕	illustrare	⊕	illustrieren	0.480	c	bn:89587v
	dedicate	⊕	dedicare	⊕	widmen	0.477	c	bn:86428v (...)
	love	⊕	amare	⊕	lieben	0.470	c	bn:90504v
	translate	⊕	tradurre	⊕	übersetzen	0.422	c	bn:89840v
	judge	⊕	giudicare	⊕	richten	0.415	c	bn:90001v
	finish	⊕	finire	⊕	beenden	0.412	c	bn:85475v
...	...		...		...	...	...	...
adj.								
↓	printed	⊕	stampato	⊕	gedruckt	0.522	c	*<no synset>*
	entertaining	⊕	divertente	⊕	unterhaltsam	0.426	c	*<no synset>*
	electronic	⊕	elettronico	⊕	elektronisch	0.466	c	bn:00102099a
	interesting	⊕	interessante	⊕	interessant	0.460	c	bn:00105276a
	famous	⊕	famoso	⊕	berühmt	0.446	c	bn:99411a
	old	⊕	vecchio	⊕	alt	0.438	c	bn:00104306a (...)
	available	⊕	disponibile	⊕	erhältlich	0.433	c	*<no synset>*
	hard	⊕	difficile	⊕	schwer	0.407	c	bn:00101358a
	open	⊕	aperto	⊕	offen	0.403	c	bn:00107879a (...)
	long	⊕	lungo	⊕	lang	0.356	c	bn:00106124a
...	...		...		...	...	...	...

- Through the SpaCy library[10], we extract and lemmatize the relevant keywords present in the gloss sentences.
- We then create a set of related words (context words) from BabelNet, which comprises the gloss keywords and terms sourced from outgoing edges.
- We analyze the context words to understand the number of relationships between different synsets and their associated lexicalizations in BabelNet.
- Finally, we compare the number of related words generated from BabelNet with those produced by our system for the same synset.

To conduct our testing, we selected a sample of 500 concepts (each associated with a given synset) that constitute polysemous words in at least one of the three languages (English, Italian, and German). The selection was done randomly to ensure a diverse and representative set of concepts. Our system generated non-empty alignments for 456 of the 500 chosen concepts. Specifically, the initial implementation of our methodology successfully discovered a total of 76,152 multilingual alignments among the 456 concepts, revealing over 80% new semantic relations compared to what is currently encoded in BabelNet across the three languages. In Table 5 we report the results of the alignments on six concepts.

As demonstrated in the Table, our system especially outperformed in retrieving new conceptual links between words - in the form of multilingual alignments- for English items, unveiling on average more than 88% new semantic relations with respect to the BabelNet database. Yet, the extracted data represent mostly unbiased, language-unbounded, and disambiguated knowledge.

Although our methodology has some limitations, such as its restriction to a small set of languages belonging to a single linguistic area and its relatively lower disambiguation performance with highly abstract and generic concepts (such as *action* (EN); *azione* (IT); *Aktion* (DE)), the results demonstrate the potential of our methodology for knowledge acquisition and suggest its possible application in constructing a novel, extensive, and cross-lingual lexical repository.

In (Grasso, Lovera Rulfi, and Di Caro 2022) a proposal for such an original lexical resource is presented. This resource is created using data collected from the previously described extraction task. The proposed model, known as "MultiAlignNet", leverages the multilingual alignments obtained through the methodology to construct a new large-scale, multilingual lexical database based on prototypical knowledge.

6. A multi-faceted semantic model

The presented knowledge model can be viewed and exploited at different levels. At the most basic level, it can be utilized as a reference for disambiguating word senses. Moreover, it can be useful for many downstream applications and studies. In this section, we briefly present an overview of such versatility.

10 https://spacy.io

Table 5
Alignments for six ambiguous concepts and percentage of unveiled *novel* relations in each language with respect to the BabelNet database. Some examples of triplets for the concept *scale-bilancia-Waage* (bn:*00069470n*) are shown in Table 3.

	00008050n	00069470n	00069470n	00062766n	00008364n	00008363n	
(en)	*libra*	*scale*	*plane*	*plane*	*bank*	*bank*	
(it)	*bilancia*	*bilancia*	*aereo*	*piano*	*banca*	*riva*	
(de)	*Waage*	*Waage*	*Flugzeug*	*Ebene*	*Bank*	*Ufer*	
triplets	**26**	**74**	**272**	**151**	**349**	**80**	
novel(en)	88,46%	87,84%	88,97%	89,40%	87,68%	91,25%	**88,9%**
novel(it)	76,92%	66,22%	75,74%	73,51%	75,64%	68,75%	**72,8%**
novel(de)	88,46%	74,32%	87,87%	84,11%	81,66%	76,25%	**82,1%**

6.1 Sense clustering

Our methodology for knowledge extraction enables the computation of proximity scores among senses, which can then be used to form clusters. These clusters can be valuable in improving the disambiguation process from a cross-lingual perspective. Consider the English word "book" as an example. This term can be assigned to several different senses in WordNet, including:

- book-1: *a written work or composition that has been published (printed on pages bound together).*
- book-2: *physical objects consisting of a number of pages bound together.*
- book-3: *the sacred writings of the Christian religions (Bible).*

By computing the intersection of the alignments among $MC^{EN,IT,DE}_{book-1}$, $MC^{EN,IT,DE}_{book-2}$ and $MC^{EN,IT,DE}_{book-3}$, we can determine the degrees of similarity between these senses, based on shared semantic information. This results in an explainable similarity score. For instance, *book-1* and *book-2* share more than a half of their $MC^{EN,IT,DE}$s, while *book-1* and *book-3* share only around 11%. Table 6 provides examples of these overlapping $MC^{EN,IT,DE}$s and shows unique semantic information related to *book-3*. This semantic information is the outcome of subtracting the common $MC^{EN,IT,DE}$s between *book-1* and *book-3* from the total alignments of *book-3*. Interestingly, the result suggests that *book-3* can be considered a type of *book-1* with the special feature of being *holy*$\oplus$*sacro*$\oplus$*Heilig*, as illustrated in the appropriate table field. Such proximity scores and similarity measures can be leveraged to improve the accuracy of disambiguation algorithms and enable the creation of more precise and meaningful word embeddings. This can have important implications for downstream applications in Natural Language Processing and other related fields.

6.2 Cross-lingual disambiguation contexts

The proposed knowledge model presents a unique approach to disambiguation compared to standard methodologies for Word Sense Disambiguation (WSD). Unlike traditional methods that rely on analyzing the context of occurrence of an ambiguous term to

Table 6
Fragment of shared $MC^{EN,IT,DE}$s among *book-1*, *book-2*, and *book-3*, and between the pair <*book-1*, *book-2*>, along with unique semantic information related to *book-3* obtained through subtraction of shared $MC^{EN,IT,DE}$s between *book-1* and *book-3*.

$MC^{EN,IT,DE}_{book-1} \cap MC^{EN,IT,DE}_{book-2} \cap MC^{EN,IT,DE}_{book-3}$
write⊕*scrivere*⊕*Schreiben*, *read*⊕*leggere*⊕*Lesen*, *text*⊕*testo*⊕*Text*, *history*⊕*storia*⊕*Geschichte*, *word*⊕*parola*⊕*Wort*
$MC^{EN,IT,DE}_{book-1} \cap MC^{EN,IT,DE}_{book-2}$
buy⊕*acquistare*⊕*kaufen*, *art*⊕*arte*⊕*Kunst*, *novel*⊕*romanzo*⊕*Roman*, *editor*⊕*editore*⊕*Herausgeber*, *write*⊕*scrivere*⊕*Schreiben*, *read*⊕*leggere*⊕*Lesen*, *text*⊕*testo*⊕*Text*, *history*⊕*storia*⊕*Geschichte*, *word*⊕*parola*⊕*Wort*
$MC^{EN,IT,DE}_{book-3} - (MC^{EN,IT,DE}_{book-1} \cap MC^{EN,IT,DE}_{book-3})$
holy⊕*sacro*⊕*Heilig*

determine its correct meaning, our model generates multilingual lexical chains that can further guide the disambiguation process. This approach is particularly useful when the context is ambiguous or when the context alone is insufficient to disambiguate the word's meaning. For example, consider the word "*wood*" in the sentence:

"*The dark* ***wood*** *hides many secrets*".

Instead of processing the underlined Bag-of-Words (BoW) representation $\{The, dark, hides, many, secrets\}$ containing "*wood*", which may not provide enough information to disambiguate the sense of the word (for example, in this context it could refer to either "*the hard fibrous lignified substance under the bark of trees*" or "*the trees and other plants in a large densely wooded area*"[11]), we can inspect an $MC^{EN,IT,DE}$ containing the word "wood" from our gathered data:

$$\begin{array}{c} \text{foresta}^{IT} \\ \oplus \\ \textit{The dark } \mathbf{wood}^{EN} \textit{ hides many secrets.} \\ \oplus \\ \text{Wald}^{DE} \end{array}$$

In this case, we can use the alignment $foresta^{IT} \oplus wood^{EN} \oplus Wald^{DE}$ to determine the correct sense of "wood". This multilingual alignment provides additional information that can be used to disambiguate the sense of the word regardless of the context in which it appears. Thus, the illustrated vertical axis may represent an additional channel that enhances the accuracy of disambiguation processes. This can be particularly use-

11 Both textual definitions are taken from WordNet.

ful in cases where context-based disambiguation methods fail to provide satisfactory results.

6.3 Concept pairs

The model presented in this work can be a powerful tool for extracting and representing semantic information between concepts. It may be seen as composed of *concept pairs*, that is, a binary set of two concepts, each identified by a specific synset. A single binary set will be composed of a concept c (a MC^k *head*) and one of the extracted concepts related to c (its MC^k). Concept pairs can be used to provide more nuanced information than simple (and ambiguous) term co-occurrences. More in detail, given a single concept c and the set of its extracted alignments within its MC^k *head*, concept pairs can be constructed by pairing c with its related MC^ks associated with existing synsets. By considering the various relationships between concepts, the model can capture a wider range of semantic knowledge. This approach may represent a unique advantage over traditional methods that rely solely on the co-occurrence of words. For instance, if we consider the concept *language* (*bn:00049910n*) and its two related MC^ks below:

word⊕*parola*⊕*Wort*
(*bn:00081546n*)

text⊕*testo*⊕*Text*
(*bn:00076732n*,*bn:00069638n*)

we can generate the following three concept pairs:

<*language:49910n*, *word:81546n*>
<*language:49910n*, *text:76732n*>
<*language:49910n*, *text:69638n*>

Each concept pair contains the original concept *language* (*bn:49910n*) paired with one of the extracted concepts *word* (*bn:81546n*), *text* (*bn:76732n*), and *text* (bn:69638n). Concept pairs could be incorporated into NLP tasks to better capture the underlying semantic relationships between concepts and improve their accuracy.

7. Conclusions and future work

In this article, we addressed the issue of knowledge encoding as a critical task in the NLP field by proposing a more inclusive approach to lexical semantics starting from the core principles of semantic prototype and linguistic representativeness. Based on this theoretical framework, we proposed an original methodology for acquiring and encoding language-unbounded, prototypical lexical knowledge through a corpus-based mechanism of multilingual alignment of semantically equivalent words. The presented model provides a cross-lingual representation of multifaceted, empirically determined conceptual links consisting of syntagmatic and paradigmatic lexical relations between words in *k* different languages. The lexical material depicted through the model is thus varied, disambiguated, and language-unbounded since the proposed methodology is meant to minimize strong language specificities and lexicographic biases. A simple implementation and experimentation over 456 concepts led to unveiling around 76K lexical-semantic alignments in three different languages (Italian, German, and English), of which more than 80% resulted as new when compared with a current state-of-the-

art resource such as BabelNet. Finally, we showed how this model suits well for a variety of applications useful for NLP tasks. Future directions include the use of more languages and large-scale runs over thousands of main concepts (Bentivogli et al. 2004; Di Caro and Ruggeri 2019; Camacho-Collados and Navigli 2017), this last being already introduced in (Grasso, Lovera Rulfi, and Di Caro 2022).

References

Apidianaki, Marianna. 2013. LIMSI : Cross-lingual word sense disambiguation using translation sense clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 178–182, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Atkins, Sue, Charles J. Fillmore, and Christopher R. Johnson. 2003. Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography - INT J LEXICOGR*, 16:251–280, 09.

Baisa, Vít, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803, Portorož (Slovenia), 23-28 May.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 94–101, Geneva, Switzerland, August.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257, 01.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270, Berkeley, California, USA, June. Association for Computational Linguistics.

Buchanan, Erin Michelle, Jessica L. Holmes, Marilee L. Teasley, and Keith A. Hutchison. 2013. English semantic word-pair norms and a searchable web portal for experimental stimulus creation. *Behavior Research Methods*, 45:746–757.

Camacho-Collados, Jose and Roberto Navigli. 2017. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, Valencia, Spain, April.

Chan, Yee Seng and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, page 1037–1042, Pittsburgh, Pennsylvania, USA, July. AAAI Press.

Devereux, Barry J., Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The cslb concept property norms. *Behavior research methods*, 46(4):1119–1127.

Di Caro, Luigi and Alice Ruggeri. 2019. Unveiling middle-level concepts through frequency trajectories and peaks analysis. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19)*, pages 1035–1042, Limassol, Cyprus, April.

Diab, Mona Talat and Philip Resnik. 2003. *Word Sense Disambiguation within a Multilingual Framework*. Ph.D. thesis, USA. AAI3115805.

Egbert, Jesse, Tove Larsson, and Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Elements in Corpus Linguistics. Cambridge University Press.

Evans, Vyvyan. 2006. *Cognitive linguistics*. Edinburgh University Press.

Fauconnier, Gilles. 1997. *Mappings in Thought and Language*. Cambridge University Press.

Fillmore, Charles J. 1977. Scenes-and-frames semantics. *Linguistic structures processing*, 59:55–88.

Grasso, Francesca, Vladimiro Lovera Rulfi, and Luigi Di Caro. 2022. Multialignet: Cross-lingual knowledge bridges between words and senses. In Oscar Corcho, Laura Hollink, Oliver Kutz, Nicolas Troquard, and Fajar J. Ekaputra, editors, *Knowledge Engineering and Knowledge Management*, pages 36–50, Cham. Springer International Publishing.

Hampton, James A. 2015. Categories, prototypes and exemplars. In *The Routledge handbook of semantics*. Routledge, pages 125–141.
Hanks, Patrick. 2004. Corpus pattern analysis. In Geoffrey Williams and Sandra Vessier, editors, *Proceedings of the 11th EURALEX International Congress*, volume 1, pages 87–98, Lorient, France, July. Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.
Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.
Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July.
Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge Applied Linguistics. Cambridge University Press.
Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 95–105, Beijing, China, July.
Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *Proceedings of the 7th International Corpus Linguistics Conference (CL 2013)*, pages 125–127, Lancaster University, UK, July.
Kecskes, Istvan, 2012. *7. Encyclopaedic knowledge and cultural models*, pages 175–198. De Gruyter Mouton, Berlin, Boston.
Kiefer, Ferenc. 1988. Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography. In T. Magay and J. Zigány, editors, *Proceedings of the 3rd EURALEX International Congress*, pages 1–10, Budapest, Hungary, September. Akadémiai Kiadó.
Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: Ten years on. *The Lexicography*, 1(1):7–36.
Kumar, Sawan, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy, July.
Lacerra, Caterina, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. Csi: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34(05), pages 8123–8130, New York, USA, February. Association for the Advancement of Artificial Intelligence.
Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.
Lefever, Els and Véronique Hoste. 2013. SemEval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
Leone, Valentina, Giovanni Siragusa, Luigi Di Caro, and Roberto Navigli. 2020. Building semantic grams of human knowledge. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2991–3000, Marseille, France, May.
McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
Moerdijk, Fons, Carole Tiberius, and Jan Niestadt. 2008. Accessing the anw dictionary. In *Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX '08)*, pages 18–24, Manchester, United Kingdom, August.
Morris, Jane and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 46–51, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.

Nelson, Douglas L., Cathy McEvoy, and Simon J. Dennis. 2000. What is free association and what does it measure? *Memory & Cognition*, 28:887–899.

Nelson, Douglas L., Cathy McEvoy, and Thomas A. Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36:402–407.

Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 14, pages 1532–43, Doha, Qatar, October.

Petricca, Paolo. 2019. *SEMANTICA. Forme, modelli, problemi*. 10.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Ruggeri, Alice, Luigi Di Caro, and Guido Boella. 2019. The role of common-sense knowledge in assessing semantic association. *Journal on Data Semantics*, 8(1):39–56.

Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34(05), pages 8758–8765, New York, USA, February. Association for the Advancement of Artificial Intelligence.

Speer, Robert, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, San Francisco, California, USA, February.

Trampuš, Mitja and Blaz Novak. 2012. Internals of an aggregated web news feed. In *Proceedings of the 15th International Multiconference on Information Society*, pages 221–224, Ljubljana, Slovenia, October.

Tulving, Endel. 1983. Elements of episodic memory.

Woods, William A. 1975. What's in a link: Foundations for semantic networks. In *Representation and understanding*. Elsevier, pages 35–82.

Zock, Michael and Chris Biemann. 2020. Comparison of different lexical resources with respect to the tip-of-the-tongue problem. *Journal of Cognitive Science*, 21(2):193–252.