Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages

Federico Martelli[1], Roberto Navigli[1], Simon Krek[2], Jelena Kallas[3], Polona Gantar[4], Svetla Koeva[5], Sanni Nimb[10], Bolette Sandford Pedersen[8], Sussi Olsen[8], Margit Langemets[3], Kristina Koppel[3], Tiiu Üksik[3], Kaja Dobrovoljc[2], Rafael-J. Ureña-Ruiz[9], José-Luis Sancho-Sánchez[9], Veronika Lipp[11], Tamás Váradi[12], András Győrffy[11], Simon László[11], Valeria Quochi[14], Monica Monachini[14], Francesca Frontini[14], Carole Tiberius[13], Rob Tempelaars[13], Rute Costa[6], Ana Salgado[6] [7], Jaka Čibej[2] and Tina Munda[2]

[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome, Italy
[2]Artificial Intelligence Laboratory, Jožef Stefan Institute, Slovenia
[3]Institute of the Estonian Language, Estonia
[4]Faculty of Arts, University of Ljubljana, Slovenia
[5]Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria
[6]NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
[7]Academia das Ciências de Lisboa, Portugal
[8]University of Copenhagen, Denmark
[9]Centro de Estudios de la Real Academia Española, Spain
[10]Society for Danish Language and Literature, Copenhagen, Denmark
[11]Hungarian Research Centre for Linguistics, Institute for Lexicology, Hungary
[12]Hungarian Research Centre for Linguistics, Institute for Language Technologies and Applied Linguistics, Hungary [13]Instituut voor de Nederlandse Taal, The Netherlands [14]Istituto di Linguistica Computazionale "A. Zampolli", Centro Nazionale delle Ricerche, Italy
Email: federico.martelli@uniroma1.it, roberto.navigli@uniroma1.it, simon.krek@ijs.si, jelena.kallas@eki.ee, apolonija.gantar@ff.uni-lj.si, svetla@dcl.bas.bg, kaja.dobrovoljc@ijs.si, lipp.veronika@nytud.hu, varadi.tamas@nytud.hu, simon.laszlo@nytud.hu, gyorffy.andras@nytud.hu, valeria.quochi@ilc.cnr.it, monica.monachini@ilc.cnr.it, francesca.frontini@ilc.cnr.it, jaka.cibej@ijs.si, tina.munda@ijs.si, bspedersen@hum.ku.dk, saolsen@hum.ku.dk, margit.langemets@eki.ee, kristina.koppel@eki.ee, tiiu.yksik@eki.ee, rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, carole.tiberius@ivdnt.org, rob.tempelaars@ivdnt.org

Abstract

Over the course of the last few years, lexicography has witnessed the burgeoning of increasingly reliable automatic approaches supporting the creation of lexicographic resources such as dictionaries, lexical knowledge bases and annotated datasets. In fact, recent achievements in the field of Natural Language Processing and particularly in Word Sense Disambiguation have widely demonstrated their effectiveness not only for the creation of lexicographic resources, but also for enabling a deeper analysis of lexical-semantic data both within and across languages. Nevertheless, we argue that the potential derived from the connections between the two fields is far from exhausted. In this work, we address a serious limitation affecting both lexicography and Word Sense Disambiguation, i.e. the lack of high-quality sense-annotated data and describe our efforts aimed at constructing a novel entirely manually annotated parallel dataset in 10 European languages. For the purposes of the present paper, we concentrate on the annotation of morpho-syntactic features. Finally, unlike many of the currently available sense-annotated datasets, we will annotate semantically by using senses derived from high-quality lexicographic repositories.

Keywords: Digital lexicography; Natural Language Processing, Computational Linguistics, Corpus Linguistics; Word Sense Disambiguation.

1. Introduction

The fields of lexicography and Word Sense Disambiguation (WSD), i.e. the area of Natural Language Processing (NLP) concerned with identifying the meaning of a word in a given context (Bevilacqua et al., 2021), are increasingly interconnected. The reasons for this are manifold. On the one hand, since the so-called statistical revolution of the late 1980s, lexicography has benefited greatly from the development and constant refinement of automatic approaches for the lexical semantic analysis of vast amounts of textual data (Johnson, 2009). On the other hand, by its very nature WSD relies heavily on the availability of wide-coverage sense repositories such as monolingual and multilingual dictionaries or lexical knowledge bases (LKBs), e.g. WordNet[1] (Miller et al., 1990) and BabelNet[2] (Navigli & Ponzetto, 2012). Furthermore, modern lexicography and WSD are inextricably tied to corpora, i.e. large collections of written text in machine-readable form. Indeed, while lexicographers analyse corpora to identify and record relevant linguistic phenomena for the purpose of creating and updating dictionaries, WSD exploits corpora in multiple ways, such as learning effective unsupervised dense representations (Devlin et al., 2019; Conneau et al., 2020), or producing training and test data to be used in supervised approaches (Vial et al., 2019; Huang et al., 2019; Bevilacqua & Navigli, 2020; Blevins & Zettlemoyer, 2020; Conia & Navigli, 2021) by annotating them in a manual, semi-automatic or fully-automatic fashion.

Interestingly, both fields suffer from the paucity of standardised manual sense-annotated data in different languages, especially low-resource ones. In fact, the majority of high-quality sense-annotated datasets focus primarily on English. This is the case, for example, with SemCor (Miller et al., 1993) and those datasets introduced in the context of the Senseval and SemEval competitions (Edmonds & Cotton, 2001; Snyder & Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2007). The few exceptions (Agirre et al., 2010; Navigli et al., 2013; Moro & Navigli, 2015) included just a limited number of instances in languages other than English. To cope with this shortcoming, several attempts have been made to bootstrap multilingual sense-annotated datasets. Pasini & Navigli (2017); Scarlini et al. (2019); Barba et al. (2020); Procopio et al. (2021) all addressed the lack of sense-annotated data in languages other than English via cross-lingual label propagation. Recently, Pasini et al. (2021) proposed XL-WSD, a large-scale multilingual evaluation framework for WSD, extending the Senseval and SemEval datasets using an automatic approach. However, despite the efforts undertaken, existing datasets are either not entirely manually curated, or they lack coverage in terms of languages, domains and parts of speech, or they rely on outdated sense inventories, which severely hampers their effectiveness.

In order to successfully address the aforementioned limitations, we have initiated the creation of a novel, manually-curated dataset, which will feature five annotation layers, i.e. tokenisation, sub-tokenisation, lemmatisation, POS tagging and Word Sense Disambiguation. The dataset will be available in 10 European languages, namely Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish. Importantly, in contrast to existing manually annotated datasets, we will annotate our dataset using high-quality sense inventories. This will enable the highest possible number of sense instances to be covered. Moreover, we will also link the annotated instances to a multilingual sense repository, namely, BabelNet, so as to allow WSD systems to use our dataset as a challenging new evaluation benchmark. In what follows, we first describe how we constructed the dataset; next, we illustrate the annotation process focusing on the first four annotation layers, and finally we describe the sense inventories which we will use to semantically annotate our dataset.

[1] https://wordnet.princeton.edu/

[2] https://babelnet.org/

Language	Tokens	Unique Lemmas	Nouns	Verbs	Adjs	Advs
Bulgarian	33,994	6,683	7,892	3,970	3,313	1,157
Danish	32,524	6,832	7,322	3,099	2,626	1,677
Dutch	34,923	6,488	7,142	3,004	2,833	1,020
English	34,228	6,297	6,716	2,946	2,818	1,079
Estonian	37,693	6,112	8,189	3,327	2,310	1,487
Hungarian	29,657	7,457	6,930	2,485	3,561	1,173
Italian	39,067	6,371	7,864	3,022	2,961	1,368
Portuguese	38,723	6,260	7,372	3,181	2,757	1,302
Slovene	31,237	6,688	7,550	2,579	3,820	1,077
Spanish	37,693	6,112	8,189	2,806	3,141	1,140

Table 1: Number of tokens, unique lemmas and open-class parts of speech.

2. Construction of the dataset

In this section, we illustrate the construction of our dataset. This process was divided into two steps: i) the automatic extraction of candidate sentences from a wide-coverage parallel corpus, and ii) the manual validation of sentences to be included in our dataset, according to specific linguistic criteria. In what follows, we detail the two phases.

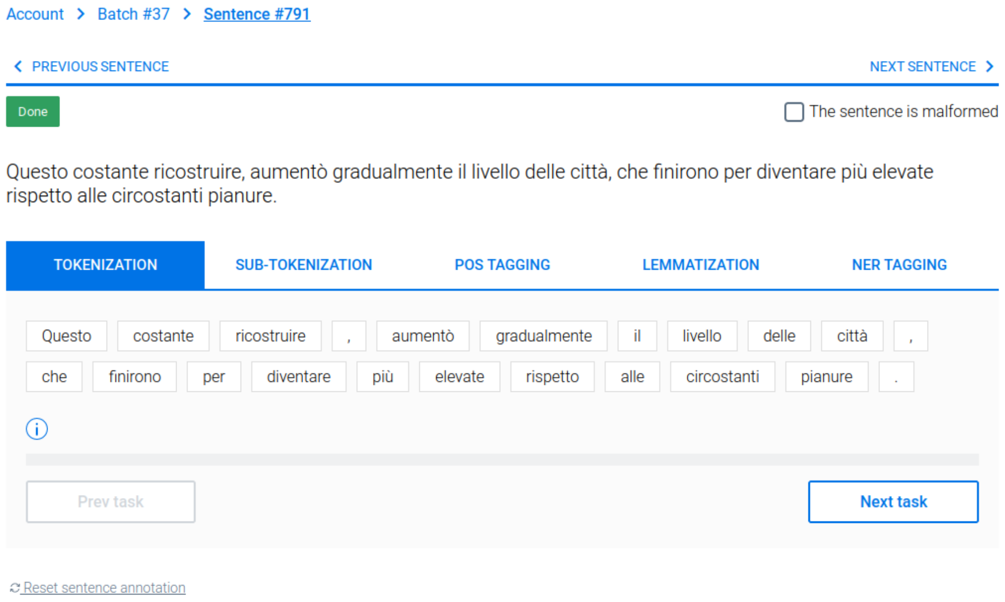
2.1 Automatic extraction of candidate sentences

First, we automatically extracted a set of sentences from WikiMatrix[3] (Schwenk et al., 2019), a large open-access collection of parallel sentences derived from Wikipedia using an automatic approach based on multilingual sentence embeddings. WikiMatrix covers 85 languages and includes 135 million parallel sentences for 1,620 language pairs, out of which 34 million are aligned with English. The corpus is divided into different files, each containing sentence pairs in a specific language combination. We performed our extraction in the following steps: i) we considered only language combinations such that the first language was English and the second was one of the other target languages; ii) we computed an overlap matrix which, given an English sentence s_e, showed the number of the selected WikiMatrix datasets which contained s_e and its corresponding translation into a target language; ii) we extracted the first 2,500 English sentences with the highest overlap across all language combinations.

2.2 Manual validation of parallel sentences

After completion of the first step, we manually validated the automatically extracted sentences according to specific formal and lexical-semantic criteria. In particular, we removed incorrect punctuation, morphological errors, notes in square brackets and etymological information typically provided in Wikipedia pages. Furthermore, in an effort to obtain a satisfying semantic coverage, we filtered out sentences which did not contain

[3] https://ai.facebook.com/blog/wikimatrix/



Questo costante ricostruire, aumentò gradualmente il livello delle città, che finirono per diventare più elevate rispetto alle circostanti pianure.

Questo | costante | ricostruire | , | aumentò | gradualmente | il | livello | delle | città | , | che | finirono | per | diventare | più | elevate | rispetto | alle | circostanti | pianure | .

Figure 1: Annotation interface used for the morpho-syntactic layers (the NER-tagging annotation was not performed at this stage).

at least 5 words, out of which at least two were polysemous. Subsequently, in order to obtain datasets in the other nine target languages, for each selected sentence in English we retrieved the corresponding WikiMatrix translation into each of the other languages. If no translation was available, we translated the English sentence manually. After the translation process, we reviewed the final dataset automatically and manually. As a result, we obtained a dataset composed of 2024 sentences for each target language.

3. Annotation

In this section, we describe the annotation process and highlight some significant linguistic peculiarities impacting on the annotation. We divided the annotation process into two phases. In the first, we focused on tokenisation, sub-tokenisation, lemmatisation and POS tagging. In the second phase we will annotate our dataset with senses derived from the specified inventories. In this paper, we will concentrate on the first phase only. In order to carry out our annotation, we used the ad hoc interface illustrated in Figure 1, developed at Babelscape[4], a Sapienza University spinoff company. In order to minimise the impact of subjectivity and ensure data consistency, we outlined specific criteria which we now detail. First, as a general guideline, we decided to follow the Universal Dependencies[5] (UD) standard for each language, so as to allow for a consistent annotation of lexical-semantic instances across languages. Importantly, we annotated both concepts and named entities. Furthermore, we normally included sub-tokenisation in almost all cases in which the token was composed of two or more distinct lemmas. As we shall see, sub-tokenisation was particularly challenging, especially when dealing with Germanic languages such as Dutch and Danish. Another challenge was posed by adjectival participles, which are derived from verbs but used as adjectives. In these cases, each annotator was required not only to consider the UD annotation standard, but also to thoroughly analyse the context in

[4] https://babelscape.com/
[5] https://universaldependencies.org/

which such instances occurred and their grammatical function in order to provide the correct tag. Table 1 reports the number of tokens, unique lemmas and the open-class part-of-speech distribution for each of the target languages.

In the following subsections, we focus on significant linguistic issues encountered during the annotation process, and provide reasons for our tagging choices.

3.1 Bulgarian

The simple and derived words, including the proper names, contractions, abbreviations and numerical expressions, were automatically annotated with the Bulgarian language processing chain (Koeva et al., 2020). This ensured the correct tokenisation, part-of-speech tagging and lemmatisation of homonymous verb particles, personal and possessive pronouns, derived numerals and proper names which were not present in the morphological dictionary. The main effort during the manual evaluation and correction was directed towards the re-annotation of multiword named entities as proper names. There are fixed multiword named entities which do not change either in terms of word order or grammar (*Yuzhna Amerika* 'South America') and semi-fixed multiword named entities which also have fixed word order but their constituents in Bulgarian can undergo certain paradigmatic changes within certain grammatical categories (for example, *Britanski muzey* 'British museum' – singular, indefinite, and *Britanskiya muzey* 'the British museum' – singular, definite). Some parts of the multiword names which can be used separately as common nouns had to be marked as proper nouns (for example, all constituents at the organization name *Evropeyski socialen fond* 'European Social Fund', etc.). The lemmas of semi-fixed multiword names in many cases were re-annotated because they differed from the lemmas of the corresponding simple words (for example, the lemmas of the words *ruskata* 'Russian' and *pravoslavna* 'orthodox' from the named entity *Ruskata pravoslavna carkva* 'Russian Orthodox Church' were changed from singular masculine to singular feminine).

3.2 Danish

The most prominent challenge in the Danish dataset was how to deal with compounds, which, as for most Germanic languages, are quite common and relatively dynamically generated, and more importantly: they are written as a single word. Our decision across all 10 languages was that conventionalized compounds found in the dictionary of the language should be kept as such, while compounds not found in the dictionary should be split into lemmas included in the dictionary, so as to enable them to be semantically tagged. For Danish we used the Danish Dictionary (DDO). When splitting compounds with a binding element, e.g. 's' in *helbredsanliggender* (health matters), we decided to keep the binding element during the subtokenisation and POS-tagging phase and to remove it in the lemmatisation phase. A further problem pertaining to compounds concerned the quite frequent phenomenon where two compounds that share a head are split and one head is left out, as in *certificeringsog revisionsmyndighed* (certification and audit authority). One possibility was to insert the head for both parts *certificeringsmyndighed og revisionsmyndighed* in the subtokenisation phase, but we decided that the head in the second part suffices for the disambiguation task and consequently we only annotated *certificerings-*'. The DDO was also used in the cases of participles used as adjectives.

Participles with adjective entries in the dictionary were annotated as such, e.g. *udstrakt* (Eng. outstretched, fig: extensive), while those that had only verb entries in the dictionary were annotated as verbs, e.g. *samlede* (Eng. lit: assembled, fig: total).

3.3 Dutch

Similarly to Danish, compounds also represented a specific challenge for Dutch. In this case, a compound was initially subtokenised if it did not occur in the Van Dale dictionary[6]. Later, this criterion was slightly relaxed and some other transparent compounds were also subtokenised, as we observed that a substantial number of compounds would not otherwise be found in the sense inventory. As in Danish, the binding element of compounds was kept in the subtokenisation phase, but removed in the lemmatisation one. Overall, 620 compounds were subtokenised in the Dutch dataset, mostly in two parts, but sometimes even in three or four parts (e.g. *laryngotracheobronchopneumonitis laryngo*; *tracheo*; *broncho* and *pneumonitis*).

An important subclass of compounds in Dutch is formed by the separable complex verbs. These are combinations of a verb and some other word. Examples are *aanvallen* 'to attack', *plaatsvinden* 'to take place'. They sometimes behave as one word (*het kan plaatsvinden* 'it can take place') and sometimes as two (*wanneer vindt het plaats?* 'when does it take place?'). Separable complex verbs are a known problem in corpus linguistics in Dutch and they presented another challenge for the annotation task. According to the UD guidelines, which are based on a lexicalist view of syntax, separable verbs should be annotated as separate words if they are written as separate words and the dependency relation should be used to identify them. After long discussions, it was decided to deviate from the UD guidelines and to consistently lemmatise separable complex verbs with the 'complex' lemma, regardless of whether the parts were separated or not. The latest version of the Alpino parser[7] also does this and lemmatises separable complex verbs with the 'complex' form, including an underscore to mark that it can occur as one word or as two, e.g. *aan_vallen.* The advantage of lemmatising with the complex verb is that the whole verb will be automatically looked up in the semantic annotation phase. This is important, as the meaning of separable complex verbs is not always compositional. Moreover, in some instances the parts of a separable complex verb are not even existing Dutch lemmas, as in the case of *aanmoedigen* 'encourage', which can be split into *aan* and *moedigen*, but where *moedigen* cannot occur on its own.

3.4 English

In the annotation of the English dataset, the scarce English-specific UD guidelines were complemented with querying the two largest manually annotated English UD treebanks – EWT (Silveira et al., 2014) and GUM (Zeldes, 2017), especially for resolving lexicon-based linguistic issues. Among others, these included the tokenisation of compounds (e.g. *cease-fire*), lemmatization of group names (e.g. *Muslims*), classification of determiner-like words (e.g. *its*), and the classification of various types of verb particle (e.g. *speed up*). Where there were discrepancies between the two treebanks, which was often the case with

[6] https://zoeken.vandale.nl/

[7] The Dutch UD corpora consist of data annotation with the Alpino annotation tools and guidelines, but do not yet include this. https://github.com/rug-compling/alpino

the under-specified lemmatisation layer, specific guidelines were drafted to consolidate the annotation of various phenomena, such as demonyms (e.g. lemma *American* of the form *American*), inflected adjectives (e.g. lemma *low* of the form *lower*) and personal pronouns (e.g. lemma *they* of the form *them*). In accordance with the general ELEXIS guidelines and the reference English treebanks, the constituents of multi-word named entities were annotated as PROPN regardless of their original POS class, with function words as an exception (e.g. *United*.PROPN *States*.PROPN *of*.ADP *America*.PROPN).

3.5 Estonian

The manual validation of the tokenisation, lemmatisation and POS tagging of the Estonian dataset generally followed the Estonian-specific UD annotation guidelines. Estonian uses 16 universal POS categories (all UD categories except PART). Regarding lemmatisation and POS tagging we relied also on the EKI Combined Dictionary[8], the biggest lexicographic database for modern Estonian compiled in the Institute of the Estonian Language. In the tokenisation phase manual correction was necessary in the case of nonconventionalised compounds (e.g. *puuja juurviljad* (fruits and vegetables)), conventionalised compounds were left as one token. For words with splitting element *Shakespeare'i* (Shakespeare's) we kept splitting elements during the subcategorisation, but removed it in the lemmatisation phase.

The most problematic was POS tagging, since Estonian UD POS tags are very different from other morphological annotators developed for Estonian (e.g. estNLTK)[9], and also from POS nomenclature used in the EKI Combined Dictionary. UD-specific parts of speech are AUX and DET. Conjunctions are also split into CCONJ and SCONJ. On the other hand, the degrees of comparison of adjectives are analysed as ADJ, while it is common for Estonian to analyse positive, comparative and superlative degrees as separate parts of speech.

According to UD annotation lemmas *olema* (to be), *ei*, *ära* (not), and modal verbs were annotated as AUX. Participles used predicatively were annotated as verbs; participles used attributively were annotated as adjectives. Abbreviations for single words were assigned the part of speech of the full form. Acronyms for proper names such as NATO were tagged as proper nouns. Foreign words were annotated as X.

3.6 Hungarian

With regard to lemmatisation and POS tagging in general we relied on the Hungarian UD guidelines[10,11] and the Magyar értelmező kéziszótár (ÉKsz. 2002) *Concise Explanatory Dictionary of Hungarian*. Regarding tokenisation, we followed the *Rules of Hungarian Orthography*, 12th edition (2015). We had to deal with the following problems in the manual correction of the result of the UD-based automatic annotation process (tokenisation, lemmatisation, POS tagging) in the Hungarian texts. First of all, the Hungarian UD POS-system is very different from the standard Hungarian POS-system

[8] http://sonaveeb.ee
[9] https://github.com/estnltk/estnltk/tree/version_1.6
[10] https://universaldependencies.org/treebanks/hu_szeged/index.html
[11] https://github.com/dlt-rilmta/panmorph

that is represented in the main explanatory dictionaries. This made the correction of the automatic POS-tagging difficult. Specific problems arose because of the lack of such categories in the UD POS-system as *igekötő* (particles or prefixes linked to verbs) and *igenév* (participles, adverbial participles and infinitives). In our explanatory dictionaries, words in the *igenév* POS-category are processed under the VERB lemmas, from which they are formed. For example, in this sentence: *A bolygót meglátogató két űreszköz...* ('The first of two spacecraft to visit the planet...'), *meglátogató* is a particle formed from the verb *meglátogat* ('to visit'). In the dictionary, there is no such lemma as *meglátogató* (because we can form participles from almost every verb). However, in the sentence this word behaves like an ADJ (attributive role), thus we have to tag it as an ADJ according to the general guidelines, even if there is no *meglátogató* ADJ in Hungarian. On the other hand, the UD POS category determiner is missing from the standard Hungarian POS-categories. (Instead, we use other categories like PRON (*egyik* ('one (of the)'), NUM (*sok* 'many')). In the annotation, we kept the DET category only for articles (*a*, *az* 'the', *egy* 'a, an'). Besides this, under the UD POS tag ADP (adposition), in Hungarian we only have postpositions. In addition, the POS tag PART is applied for only two words: *meg* and *utol.* Regarding tokenisation and lemmatisation, we had to deal with general, non-language specific problems: the multi-word proper names had to be analysed as a whole, despite the fact that they might have contained common noun elements, too. Another problematic case was presented by the ellipsis in complex compounds, in which the lemmatisation depends on whether we take the missing words into consideration, or not. In Hungarian, an agglutinative language, we also had such problems as suffixes attached to symbols (e.g. %-*át*), and suffixes after quotation marks (e.g. "xyz"-*t*). Also, the orthographic rules influence tokenisation: *Schmidt-távcső* ('Schmidt telescope') is a compound lemma (one token), a NOUN, thus, the PROPN-element is "lost" in it. (Analogy: *Kossuth-szakáll* 'a type of beard which was made famous by Lajos Kossuth' – not a PROPN).

3.7 Italian

The manual correction/validation of the morphosyntactic layers of the Italian dataset generally followed the Italian-specific UD annotation guidelines[12] and the praxis established in the Italian treebanks[13]. When clashes with project-level indications arose, it was decided to adhere to the UD praxis, with a few exceptions as follows: a) abbreviations are treated as single words that may contain punctuation (e.g. *U.S.A.*, *UE*) except when they indicate units of measure, in this case they are annotated as SYM as in the rest of the datasets; and b) foreign words are annotated as X in titles and long expressions (i.e. when they are incidentals). As for POS annotation, base infinitives used as nouns and participles used predicatively are annotated as verbs, even when the subject is implied; participles used attributively are annotated as adjectives instead. Possessive adjectives are always tagged as determiners, while predeterminers and quantifiers are tagged as such if no other determiner is present, adjective otherwise. As for POS-tags, it is worth noting that AUX is also used for copulas, so that the verb *essere* "to be" is almost always an AUX. Subtokenisation in Italian UD is required in the following cases: 1) complex prepositions (i.e. combined/fused with the definite article, e.g. *nella* "in the.fem", *del* "of the.masc"); and 2) verbal forms with enclitic pronouns (e.g. *dammelo* "give-to me-it", *mangiandolo*

[12] https://universaldependencies.org/it/index.html

[13] i.e. https://universaldependencies.org/treebanks/it_isdt/index.html, https://universaldependencies.org/treebanks/it_partut/index.html

"eating-it"). Given that they are quite frequent in training data, manual correction was not often required for these aspects. As for lemmatisation, articles and pronouns were lemmatised with their base form (i.e. singular masculine); adjectives with the positive, singular, masculine forms, except for irregular comparative and superlative forms.

Finally, regarding the incidence of manual corrections needed, lemmatisation required a considerable effort, as the automatic lemmas assigned were often wrong, especially for homographs and irregular and infrequent words.

3.8 Portuguese

One of the major challenges in annotating the Portuguese dataset was presented by lemmatisation in a dictionary that did not always abide by the same annotation criteria applied to corpora. We decided to always annotate the lemma as being the canonical form during the lemmatisation process, ignoring some of the lexical items identified that occur as a headword in a dictionary. For instance, the personal pronoun *ela* (she), the Portuguese feminine form of *ele* (he), is registered as an entry in the Portuguese dictionary, and the recorded lemma is the canonical form *ele*. The option we decided on guarantees better data consistency and coherency. In dictionaries, cases of this type often turn out to be cross-referenced to the canonical form, e.g., the definite article *a* [the Portuguese feminine form of 'the'] is a cross-reference to *o* [the Portuguese masculine form of 'the'], which strengthens our decision.

Another decision we took concerned the forms corresponding to degrees of adjectives and adverbs. Although in the Lisbon Academy of Sciences dictionary we find comparative and superlative forms as headwords, e.g., *pior* (worse; worst), we considered the positive form as a lemma according to Universal Dependencies recommendations. Generally, for the part-of-speech tagging, we used the Universal Dependencies (UD) in its current version 2.7 (Zeman et al., 2020). Nevertheless, we did not adhere to the UD criterion for abbreviations. Lexical items such as *km* (kilometre) and *m* (meter) were tagged as abbreviations as previously agreed by all ELEXIS team members, rather than as nouns, as UD suggests. It is important to note that we labelled some past participles as adjectives rather than as verbs when they served an adjectival function in the analysed sentences.

As for the subtokenisation, contractions were broken into smaller units, for example, *da* (*de* + *a*) [preposition *de* (of) + the feminine article form *a* (the)]. However, in the case of *desde* (since), which is a contracted form (< prep. Latin *de* + *ex*), we preferred instead to keep it as a preposition, as recognised by Portuguese grammar and dictionaries.

3.9 Slovene

The Slovene dataset was automatically tokenised, lemmatised and tagged with the CLASS LA tagger (Ljubešić & Dobrovoljc, 2019), which was developed for processing South Slavic languages. The tagger proved to be a highly accurate tool, although some corrections were needed.

For Slovene, two POS tagsets are generally used, the default JOS (Erjavec et al., 2010)/ Multext-East system (Erjavec, 2017), and UD (Dobrovoljc et al., 2017). Taggers usually struggle with two major differences between the systems. One difference lies in the distinction between the categories AUX and VERB in case of the omnipresent verb *biti* ('to be'). In the UD system, the AUX category is assigned when 'to be' is used as an auxiliary or a linking verb (e.g. *Večina prebivalcev je* AUX *katolikov.* The majority of the population are AUX Catholic.), and the category VERB when it is used as a lexical verb (*Njihov glavni štab je* VERB *v Tel Avivu.* Their headquarters are VERB in Tel Aviv.). In real life, the distinction between these is not always clear-cut; however, to solve the dilemmas, we consulted the detailed UD-POS tagging guidelines for Slovene (*ibid.*).

The other major difference is the use of categories DET vs PRON. In UD, the DET category is assigned to pronouns when used as modifiers in nominal (or other) phrases, and PRON when they are used as heads. Other notable issues include the use of CCONJ vs ADV (*Ali* ADV *so te razlike neposredni vzrok za debelost ali* CCONJ *pa njena posledica, je še odprto vprašanje.* Whether CCONJ these differences are the direct cause or CCONJ the result of obesity has yet to be determined unequivocally.); ADP vs ADV (*Sklepali so, da je okoli* ADP *Urana sistem obročev.* They concluded that there must be a ring system around ADP Uranus. vs *V naravi povprečno živi okoli* ADV *20 let.* The life expectancy in the wild is approximately ADV 20 years.).

In order to obtain as many content words as possible, such words being the only ones considered in the lexical-annotation phase, components of named entities that were not proper nouns were assigned the part of speech they belong to in their simple, common sense (e.g. *Evropska* ADJ *unija* NOUN; European ADJ Union NOUN). This decision is in line with the Slovene UD guidelines, but contrary to the practice of most of the project participants. As for lemmatisation, the preposition *s* ('with') was oftentimes automatically lemmatised as *biti* ('to be'), and prepositions, when occupying the first place in a sentence, were lemmatised with the capital letter, all of which was manually corrected. There were no errors in tokenisation.

3.10 Spanish

The main revision points on the annotated Spanish dataset were the lemmatisation of infrequent or rare words, verb infinitive lemmas with adjectival tags and non-toponym non-anthroponym PROPN sequences readjusting into (chiefly postmodified) common noun phrases, e.g. *Tribunal Supremo* "Supreme Court". Lemmatisation followed the standard practice of Spanish linguistics: infinitive for verbs and masculine singular form for other inflectional elements (N, ADJ, PRON, DET, but not for PART), even when some of their forms were dictionary entries, usually for alphabetical reasons or retrievability purposes.

Some functional tags also needed correction in traditional fuzzy zones of Spanish syntax such as NOUN-ADJ triplets and DET-PRON subsystems, correlative structures (e.g. *tan... como "as... as"*) and, very occasionally, complement-relative clausal misanalyses of *que* "who, that".

Tokenisation followed UD guidelines and the only subtokenised elements were verbal forms with oblique pronouns like *matarlo* "kill him". As a rule, general complex elements such as multiple verbs (compound tenses, aspectual or catenative structures and the like), comitatives and the only two amalgamated ADP-DET remnants in Spanish (*al, del*) are kept exactly as tokenised – the former split and the latter two groups not subtokenised.

Language	**Resource**
Bulgarian	Dictionary of Modern Bulgarian
Danish	DanNet (The Danish WordNet)
Dutch	Open Dutch WordNet
English	English WordNet
Estonian	EKI Combined Dictionary
Hungarian	The Explanatory Dictionary of the Hungarian Language
Italian	PAROLE-SIMPLE-CLIPS + ItalWordNet
Portuguese	Dictionary of the Lisbon Academy of Sciences
Slovene	sloWNet
Spanish	Spanish Wiktionary

Table 2: Sense inventories

4. Sense inventories

We now describe the sense inventories which we will use to annotate our dataset semantically, as shown in Table 2. Importantly, during the semantic annotation validators will be able to improve the coverage and quality of the specified sense inventories, for instance, by adding new entries or improving already existing ones.

4.1 Bulgarian

The Dictionary of Modern Bulgarian (DMB, *Rechnik na savremenniya balgarski knizhoven ezik*) was published in three volumes between 1955 and 1959 by the Bulgarian Academy of Sciences. In addition to the general vocabulary the dictionary includes some obsolete words, words gradually moving into the passive vocabulary, and foreign words which are widely used in modern Bulgarian. Each entry is structured in a specific way according to the part of speech of the headword and it represents the major senses accompanied with quotations. The headword is followed by a forms section, a grammar section, a stylistic section and an etymology (where relevant). An entry may also include compounds, phrases, and derivatives (secondary lemmas) based on the headword. Today, the dictionary is in the process of its first major revision. The update is revising and extending the DMB, adjusting the vocabulary to cover the missing senses from the ELEXIS *multilingual parallel sense-annotated dataset*, to label some senses as obsolete, to include some new borrowings in the language, and to replace the obsolete quotations. As of March 2021 the dictionary covers 60,744 headwords, 68,387 lemmas and secondary lemmas, 78,569 sense definitions and 80,520 quotations coming mainly from classic literature and periodicals.

4.2 Danish

The Danish sense inventory applied for the annotation task consists of by the Danish wordnet, DanNet (Pedersen et al., 2009). DanNet currently contains 70,000 synsets corresponding roughly to the same number of word senses, covering nouns, verbs and adjectives. The wordnet follows the Princeton WordNet standard, but is compiled semi-automatically from a Danish source, namely The Danish Dictionary (DDO), and linked to the senses in the dictionary (Pedersen et al., 2009). Approx 10,000 of the synsets are also linked to Princeton WordNet (Pedersen et al., 2019). DanNet is currently being extended to cover a broader number of word senses (Nimb et al., 2021), still in essence relying on the sense inventory of DDO as a basis, but aiming towards partly clustering very subtle meaning distinctions inherited from the source (Pedersen et al., 2018). DanNet was chosen for the annotation task first of all because it allows us to publish the annotation sense inventory as open source, but also because we want to test the lexical coverage as well as the operability of the wordnet for such a task. Based on the feedback and results, missing lemmas and senses will subsequently be added to the wordnet and further integrated into a future Danish language resource for AI purposes to be developed in COR (The Central Word Register for Danish)[14], a collaborative project between the Society for Danish Language and Literature, the Danish Language Council, Centre for Language Technology at UCPH and the Danish Agency for Digitisation.

4.3 Dutch

Open Dutch WordNet is a Dutch lexical semantic database. It was created by removing the proprietary content from Cornetto[15]. A large portion of the Cornetto database originated from the commercial publisher Van Dale[16] preventing it from being distributed as open source. In order to create Open Dutch WordNet, all the synsets and relations from WordNet 3.0 were used as a basis and existing equivalence relations between Cornetto synsets and WordNet synsets were exploited in order to replace WordNet synonyms by Dutch synonyms. Concepts that were not matched through hyperonym relations to the WordNet hierarchy were added, as well as manually created semantic relations from Cornetto. The synonyms, concepts and relations were limited to those on which there were no copyright claims. In addition, the inter-language links in various external resources were used to add synonyms to the resource (Postma et al., 2016).

4.4 English

The English WordNet[17] is an open-source derivation from Princeton WordNet (Miller, 1995), a widely used lexical network of the English language grouping words into synsets and linking them according to different semantic relations between them. In its second release, the English WordNet 2020 (McCrae et al., 2020) introduced a substantial number of changes compared to the original database, including the integration of contributions from other projects, such as Colloquial WordNet (McCrae et al., 2017), enWordNet (Rudnicka et al., 2015) and Open Multilingual WordNet (Bond & Paik, 2012). This resulted in the introduction of several new manually-validated synsets (120,054 in total), lemmas (163,079), senses (211,864) and definitions (120,059), as well as the development of clear guidelines for future community-driven additions to the database, which is planned to be released annually.

[14] https://cst.ku.dk/english/projects/the-central-word-register-for-danish-cor/

[15] http://www2.let.vu.nl/oz/cltl/cornetto

[16] https://www.vandale.nl/

[17] https://en-word.net/

4.5 Estonian

EKI Combined Dictionary[18] is the biggest lexicographic database of modern Estonian compiled in the Institute of the Estonian Language. The current description of Estonian headwords in Ekilex includes definitions, semantic types, parts of speech, inflected forms, collocations, government patterns, semantic relations, related words, etymology, usage examples, and translations. As of April 2021, Ekilex contains about 160,000 words and phrases in Estonian. For this task's development, a total of 7,044 Estonian lemmas and 14,870 senses were extracted from Ekilex. Ekilex allows the annotation sense inventory to be published as open source.

4.6 Hungarian

The Explanatory Dictionary of the Hungarian Language (*A magyar nyelv értelmező szótára*, abbr. ÉrtSz.) was compiled in the Research Institute for Linguistics, Hungarian Academy of Sciences in seven volumes between 1959 and 1962. ÉrtSz. covers Hungarian literary language of the 19th century, as well as the written and spoken standard Hungarian of the first half of the 20th century, with a total of 60,000 entries. The main source of input was a corpus of about six million examples collected since the end of the 19th century. Entries included pronunciation (where it differed from what could be expected on the basis of spelling) and an aid to the hyphenation of compound words. Each sense unit is illustrated by a few examples: citations from the classical Hungarian literature and example sentences created by the lexicographers. In terms of the fine sense discrimination and sophisticated sense definitions, it stands out from the genre of a desk dictionary and is closer in its ambitions to unabridged dictionaries, particularly as regards the treatment of function words and detailed treatment of verb senses. This is one of the best used dictionaries from a professional point of view, but its vocabulary and the examples are old-fashioned.

4.7 Italian

The Italian Sense Inventory was produced by combining two existing openly available lexical resources, namely PAROLE SIMPLE CLIPS (PSC)[19] and ItalWordNet (IWN)[20]. PSC, developed within two subsequent European projects PAROLE and SIMPLE, is a large lexical database for the Italian language. In the semantic layer, the main basic blocks are semantic units, Usems, which are provided with definitions and examples, and linked to the SIMPLE Ontology and also to other Usems through a rich set of semantic relations (Bel et al., 2000). ItalWordNet (IWN) is a lexical semantic database for the Italian language started within the context of the EuroWordNet project and then subsequently enlarged and refined within national projects until 2012. It is mapped and linked to the Princeton WordNet – thus also indirectly, to BabelNet – and is also available in the Open Multilingual Wordnet format (Quochi et al., to appear). The two resources have been partially aligned, so that a subset of IWN synsets are linked to PSC corresponding Usems. In order to produce the current sense inventory, the two resources were queried for all the target lemmas present in the Italian dataset and a list of corresponding Usems

[18] http://sonaveeb.ee
[19] http://hdl.handle.net/20.500.11752/ILC-88
[20] http://hdl.handle.net/20.500.11752/ILC-62

from PSC and IWN synsets were retrieved together with their definitions, examples and original IDs. Where a mapping between the two resources was available, a unique sense was produced, merging the two definitions into a single one. The resulting sense inventory contains 4,424 lemmas for a total number of 11,298 senses.

4.8 Portuguese

The *Dicionário da Língua Portuguesa* (DLP) is a scholarly dictionary of the Portuguese language being developed by the Lisbon Academy of Sciences. DLP is a retro-digitised dictionary created by converting the *Dicionário da Lingua Portuguesa Contemporânea*, last published in 2001. Currently, the DLP is being prepared under the supervision of the Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP) in collaboration with researchers and invited collaborators. Between 2015 and 2016, some preparatory work for the Portuguese Academy digital dictionary was performed through the ILLLP, and a database was developed by a team working in NLP at the University of Minho (Simões et al., 2016), which now includes IPCA and NOVA CLUNL (Salgado et al., 2019). This project is supported by a Community Support Fund – Fundo de Apoio à Comunidade (FAC) – by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia. For the development of this task, a total of 4,031 lemmas (lemma, part of speech, and definitions) were extracted from DLP.

4.9 Slovene

As a Slovene sense inventory, we used the current version of the Slovene wordnet – sloWNet 3.1 (Fišer & Sagot, 2015). This is an open-source lexical database containing the complete Princeton WordNet 3.0 and 71,803 Slovene literals, 33,546 of which were manually validated. The literals were inserted automatically from several existing language resources, comprising two bilingual dictionaries, a few domain-specific resources, parallel corpora, as well as Wikipedia. The 4,919 content word lemmas appearing in the dataset were validated and corrected, if necessary, during the WSD annotation process.

4.10 Spanish

To come up with a freely distributable dataset, the Spanish lexical fragment of Wiktionary[21] was chosen to tag Spanish texts. Wiktionary is a multilingual free dictionary, being written collaboratively on the web. A dump as of late 2020 was filtered to sort out non-semantic information (etymology, morphology, pronunciation, etc.) and about 92,000 lemmas with more than 140,000 senses were kept. Wiktionary has been shown (Ahmadi et al., 2020) to exhibit a great deal of overlap with the reference Spanish dictionary (Real Academia Española & Asociación de Academias de la Lengua Española, 2014), so standard coverage is envisaged.

5. Conclusions

In this work, we addressed a major shortcoming affecting both lexicography and Word Sense Disambiguation, namely the paucity of manual sense-annotated data. We

[21] https://es.wiktionary.org/wiki/Wikcionario:Portada

described how we plan to design a novel manually curated dataset available in 10 European languages, i.e. Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish, focusing on the morpho-syntactic annotation layers. We have now finalised the annotation of the morpho-syntactic layers and, as next step, we will annotate our dataset with senses derived from the aforementioned high-quality sense inventories. We argue that, thanks to our dataset, both scientific communities will be provided with a very effective resource which, on the one hand, will enable lexicographic phenomena to be investigated both within and across languages, and on the other hand, can be used as a new evaluation benchmark for WSD systems.

Acknowledgments

The authors gratefully acknowledge the support of the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme.

6. References

Agirre, E., De Lacalle, O.L., Fellbaum, C., Hsieh, S.K., Tesconi, M., Monachini, M., Vossen, P. & Segers, R. (2010). SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th international workshop on semantic evaluation*. pp. 75–80.

Ahmadi, S., McCrae, J.P., Nimb, S., Khan, F., Monachini, M., Pedersen, B.S., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgárd, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Moshe, Y.B., Rudich, M., Ahmad, R.A., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J., Ureña-Ruiz, R., Zamorano, J.P., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stankovic, R., Perdih, A. & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. European Language Resources Association, pp. 3232–3242. URL https://www.aclweb.org/anthology/2020.lrec-1.395/.

Barba, E., Procopio, L., Campolungo, N., Pasini, T. & Navigli, R. (2020). MuLaN: Multilingual Label propagatioN for word sense disambiguation. In *Proc. of IJCAI*. pp. 3837–3844.

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. & Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/61.pdf.

Bevilacqua, M. & Navigli, R. (2020). Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 2854–2864.

Bevilacqua, M., Pasini, T., Raganato, A. & Navigli, R. (2021). Recent Trends in Word Sense Disambiguation: A Survey. In *Proc. of IJCAI*.

Blevins, T. & Zettlemoyer, L. (2020). Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 1006–1017.

Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4), p. 5.

Conia, S. & Navigli, R. (2021). Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. In *Proceedings of the EACL*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 8440–8451.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186. URL https://doi.org/10.18653/v1/n19-1423.

Dobrovoljc, K., Erjavec, T. & Krek, S. (2017). The universal dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. pp. 33–38.

Edmonds, P. & Cotton, S. (2001). Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. pp. 1–5.

Erjavec, T. (2017). MULTEXT-East. In *Handbook of Linguistic Annotation*. Springer, pp. 441–462.

Erjavec, T., Fiser, D., Krek, S. & Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *LREC*. Citeseer.

Fišer, D. & Sagot, B. (2015). Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3), pp. 601–635.

Huang, L., Sun, C., Qiu, X. & Huang, X.J. (2019). GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pp. 3500–3505.

Johnson, M. (2009). How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* pp. 3–11.

Koeva, S., Obreshkov, N. & Yalamov, M. (2020). Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In *Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association*. pp. 6988–6994.

Ljubešić, N. & Dobrovoljc, K. (2019). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing*. pp. 29–34.

McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Marseille, France: The European Language Resources Association (ELRA), pp. 14–19. URL https://www.aclweb.org/anthology/2020.mmw-1.3.

McCrae, J.P., Wood, I.D. & Hicks, A. (2017). The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *LDK*.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D. & Miller, K. (1990). Introduction to WordNet: an Online Lexical Database. *International Journal of Lexicography*, 3(4).

Miller, G.A., Leacock, C., Tengi, R. & Bunker, R.T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Moro, A. & Navigli, R. (2015). SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. pp. 288–297.

Navigli, R., Jurgens, D. & Vannella, D. (2013). SemEval-2013 task 12: Multilingual 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)*. pp. 222–231.

Navigli, R., Litkowski, K.C. & Hargraves, O. (2007). SemEval-2007 task 07: Coarse-Grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. pp. 30–35.

Navigli, R. & Ponzetto, S.P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193, pp. 217–250.

Nimb, S., Pedersen, B.S. & Olsen, S. (2021). DanNet2: Extending the coverage of adjectives in DanNet based on thesaurus data. In *Proceedings of the 11th Global Wordnet Conference*. pp. 267–272.

Ogilvie, S. (2020). *The Cambridge Companion to English Dictionaries*. Cambridge University Press.

Pasini, T. & Navigli, R. (2017). Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 78–88.

Pasini, T., Raganato, A. & Navigli, R. (2021). XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In *Proc. of AAAI*.

Pedersen, B.S., Aguirrezabal Zabaleta, M., Nimb, S., Olsen, S. & Rørmann, I. (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In *Proceedings of Global WordNet Conference 2018 Singapore: Global WordNet Association*.

Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L. & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3), pp. 269–299.

Pedersen, B.S., Nimb, S., Olsen, I.R. & Olsen, S. (2019). Linking DanNet with Princeton WordNet. In *Global WordNet 2019 Proceedings, Wroclaw, Poland*.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A. & Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*. Bucharest, Romania.

Pradhan, S., Loper, E., Dligach, D. & Palmer, M. (2007). SemEval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. pp. 87–92.

Procopio, L., Barba, E., Martelli, F. & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *Proc. of IJCAI*. Online.

Quochi, V., Bartolini, R. & Monachini, M. (to appear). ItalWordNet goes open. *Special Issue on Linking, Integrating and Extending Wordnets. Linguistic Issues in LanguageTechnology. Linguistic Issues in Language Technology. LiLT*, 10(4).

Real Academia Española & Asociación de Academias de la Lengua Española (2014). *Diccionario de la lengua española*. Espasa Calpe, vigesimotercera edición edition.

Rudnicka, E.K., Witkowski, W. & Kaliński, M. (2015). Towards the methodology for extending Princeton WordNet. *Cognitive Studies/ Études cognitives*, (15), pp. 335–351.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. pp. 417–433.

Scarlini, B., Pasini, T. & Navigli, R. (2019). Just “OneSeC” for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 699–709.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H. & Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Silveira, N., Dozat, T., de Marneffe, M.C., Bowman, S., Connor, M., Bauer, J. & Manning, C.D. (2014). A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Simões, A., Almeida, J.J. & Salgado, A. (2016). Building a Dictionary using XML Technology. In M. Mernik, J.P. Leal & H.G. Oliveira (eds.) *5th Symposium on Languages, Applications and Technologies (SLATE)*, volume 51 of *OASIcs*. Germany: Schloss Dagstuhl, pp. 14:1–14:8.

Snyder, B. & Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. pp. 41–43.

Taghipour, K. & Ng, H.T. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*. pp. 338–344.

Vial, L., Lecouteux, B. & Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*.

Yuan, D., Richardson, J., Doherty, R., Evans, C. & Altendorf, E. (2016). Semi-supervised Word Sense Disambiguation with Neural Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp. 1374–1385.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3), pp. 581–612.

Zeman et al. (2020). *Universal Dependencies 2.7*. URL http://hdl.handle.net/11234/1-3424. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.