Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3008–3015
Marseille, 11–16 May 2020

A Parallel WordNet for English, Swedish and Bulgarian

Krasimir Angelov
Chalmers University, University of Gothenburg
Gothenburg, Sweden
krasimir@chalmers.se

Abstract

We present the parallel creation of a WordNet resource for Swedish and Bulgarian which is tightly aligned with the Princeton WordNet. The alignment is not only on the synset level, but also on word level, by matching words with their closest translations in each language. We argue that the tighter alignment is essential in machine translation and natural language generation. About one-fifth of the lexical entries are also linked to the corresponding Wikipedia articles. In addition to the traditional semantic relations in WordNet, we also integrate morphological and morpho-syntactic information. The resource comes with a corpus where examples from Princeton WordNet are translated to Swedish and Bulgarian. The examples are aligned on word and phrase level. The new resource is open-source and in its development we used only existing open-source resources.

Keywords: WordNet, Translation, Morphology, Sense-annotated Corpora, Wikipedia

1. Introduction

WordNet (Fellbaum, 1998) started as an effort to enumerate the senses of all English words, and to organize those senses into a network of interrelated senses. Soon after, the effort was taken into other languages. First the EuroWordNet (Bloksma et al., 1996) for Dutch, Italian, Spanish, French, German, Czech and Estonian, and later the BalkaNet (Stamou et al., 2019) for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish. There is also a multitude of projects for other languages. Most publicly available WordNet resources are aggregated by the Open Multilingual WordNet project (Bond and Paik, 2012).

Initially not all of the resources were released with free access, but luckily some of them were made free later or are being replaced with alternative free resources. Our contribution is a free WordNet-like resource for Swedish and Bulgarian. The availability of prior resources is discussed in Section 2.

We have, in addition, the goal that the new resource should be usable for applications in GF (Grammatical Framework; Ranta (2011)). GF is a programming language for natural language applications, whose primary focus is on multilinguality. Any GF application makes use of a grammar and a lexicon. It is quite common for an application to be available in five to ten languages simultaneously. Therefore, to minimize the effort for new applications, the framework offers a Resource Grammars Library (RGL) for many languages (Ranta, 2009). The lexicon, however, is typically built from scratch, since it must be highly multilingual and application specific. These kinds of lexicons are hard to get. For every language they must contain a full-form inflection table, and the words that are translations of each other must be linked together through an interlingual index.

We see the new resource as a lexical library which must be compatible with the already existing resource grammars. From there, the different applications can build specialized lexicons for different purposes. In order for that to work, just another WordNet would not be enough.

To start with, in addition to synonyms and semantic relations, the lexicon must also contain translations. Most existing resources are aligned on synset level with Princeton WordNet, but as we will show later, synset alignment is not the same as building a translation dictionary. For that reason, similar to FinnWordNet (Lindén and Carlson, 2010), we preserve both the semantic relations as well as the translation equivalence. The later is better seen as yet another relation between words in different languages.

The last piece is that, since GF is used for both parsing and generation, it is essential to also represent the morphology and certain morpho-syntactic features which are also not part of the traditional WordNet.

In the process we found it helpful to use the examples in Princeton WordNet as a guide to choose the best translations for Swedish and Bulgarian. We then took one step further and parsed all examples with the GF's grammar for English. From the GF analysis of the examples, it is possible to automatically generate translations for Swedish and Bulgarian. Their correctness, of course, depends on the correctness of the syntactic analysis, and even more on the correct choice of a word sense. When we work with the examples, whenever necessary, we first correct the automatic syntactic and semantic analysis of the statistical parser (Angelov and Ljunglöf, 2014), then we make sure that the right Swedish and Bulgarian word choices are used in order to generate reasonable translations.

While looking for translations, as one of the possible sources, we used the titles of related articles from Wikipedia. We retained the links to the original articles, which now can be accessed via our search interface.

The focus of the current report is on English, Swedish and Bulgarian, but we have actually bootstrapped similar representations for 11 other languages: Catalan, Chinese, Dutch, Estonian, Finnish, Italian, Portuguese, Slovenian, Spanish, Thai and Turkish. This was done by using existing WordNets and the automatic translation alignment method presented in Angelov and Lobanov (2016). The status of these languages is reported in Section 4.

In the next section we will first review the available resources for Swedish and Bulgarian, and after that in Section 3we will go through the different steps in building the

data. The article concludes with a report on the current status.

2. Related Work

The first WordNet for Bulgarian was built in the BalkaNet project (Koeva and Genov, 2004). However, to this date the lexicon is not freely available. For us it is important to be able to share the new resource with the community, and therefore we insisted on using only free resources. A free core-WordNet for Bulgarian was made available in the BulTreeBank Wordnet (Simov and Osenova, 2010), but unfortunately its size is rather small - 8 936 senses. On the other hand, it has very good quality, so when we had to choose a translation for a word in English, we first looked for a corresponding synset in the BulTreeBank Wordnet. It often happened, however, that the Bulgarian synset did not contain all possible translations. All existing synsets from the BulTreeBank Wordnet are integrated in our resource, but when necessary we also made our own extensions.

Another handy resource is the English-Bulgarian translation dictionary in Angelov (2014). The dictionary is compatible with GF, and contains morphology and English-Bulgarian translations. However, the translations are not sense annotated. We used that dictionary to extract the morphology and to bootstrap the translations.

The situation with Swedish is similar to Bulgarian. There is one free resource, SALDO-WordNet (Borin et al., 2013) with 6 904 senses, and the larger Svenskt OrdNät (Viberg et al., 2002) with 28 046 senses. Svenskt OrdNät was closed source until very recently when it was made available through the Swedish Language Bank (Språkbanken). All of SALDO-WordNet, plus all nouns and some of the verbs in Svenskt OrdNät are part of our lexicon too. Integration of the rest of Svenskt OrdNät is still ongoing.

For Swedish, there is also the SALDO lexicon (Borin et al., 2008), which contains complete morphology, as well as semantic relations. We used the SALDO morphology, but not the semantics since it is very different from the one in WordNet.

The last resource is the Folkets Lexikon (Kann and Hollman, 2011) which contains English-Swedish translations collected by crowd sourcing.

Finally, for both languages we also used translations from PanLex (Kamholz et al., 2014) and from titles of Wikipedia articles.

3. A WordNet in GF

The goal of compatibility with GF predetermines the data representation. In this section we will review the different choices which we made for each component in the data. We will also summarize how the data was collected and verified.

3.1. Synsets and Translations

A synset in the GF WordNet is a matrix as in Figure 1, where the rows of the matrix represent the translation equivalence, while the columns represent the synsets of the different languages. Each row is also labeled with an abstract identifier which can be used as a variable to access the tuple of translations from a GF grammar. The general convention for the identifiers is to use the English lemma plus a part of speech tag. If that does not disambiguate the meaning then we add the sense number from Princeton WordNet to the identifier.

Note that the same identifier also aligns together the translation equivalents. The raw Princeton WordNet synset is an unordered collection of words such as:

```
(08094856-n) family, household, house,
             home, menage
        (a social unit living together)
```

which is then aligned with a synset in another language, e.g.

```
(08094856-n) hus, hushåll, familj, hem
        (a social unit living together)
```

It is obvious that such a synset-level alignment does not capture the most pragmatic translations between words. The word "family" is almost always translated as "familj" in Swedish, even if it is possible to replace it with "hus" (house) for that particular sense.

Another case showing that a tighter alignment is better is synsets which contain words that are synonyms but are used in very different domains. For instance we have:

```
(12654755-n) apple, orchard apple tree,
             Malus pumila
    (native Eurasian tree widely
     cultivated in many varieties for
     its firm rounded edible fruits)
```

Here the Latin name "Malus pumila" is used in botany, and is inappropriate as an everyday term for an apple tree.

We made the translation alignment as tight as possible but it is not always a one to one relation. For instance, as it can be seen on Figure 1, in Spanish and Portuguese one and the same word is reused for different abstract identifiers. It is also possible that for some languages it is not possible to find an appropriate translation, and then just we leave a gap. Finally, when there are spelling variations and there are no obvious semantic or pragmatic differences, then we assign all variants to a single abstract identifier.

It is also worth noting that a single identifier like `family_1_N` could be reused across synsets if the same tuple of translations also fits another sense. For instance, the above example also matches the sense:

```
((biology) a taxonomic group
         containing one or more genera)
```

This is useful since the abstract identifiers become more coarse-grained sense indicators. That ought to make statistical sense disambiguation on the level of abstract identifiers more robust.

3.2. Data Collection

We bootstrapped the WordNet synsets by using a translation dictionary. For Bulgarian we started from the English-Bulgarian dictionary for GF (Angelov, 2014), and for English-Swedish we used a similar dictionary created from Folkets Lexikon and SALDO. These lexicons contain translations pairs and provide full-form inflection tables for each

Abstract	Bulgarian	English	Finnish	Portuguese	Slovenian	Spanish	Swedish
household_N	домакинство	household	kotitalous	casa	gospodinjstvo	casa	hushåll
house_10_N	къща	house	talous	casa	hiša	casa	hus
home_8_N	дом	home	perhe	lar	dom	hogar	hem
family_1_N	семейство	family	suku	família	družina	familia	familj

Figure 1: A snapshot of a synset from the search interface for GF WordNet

language. The downside, however, is that there are no sense annotations and links to WordNet. Most English words are translated to each of the two target languages but it is not known in what sense the translation is valid. Fortunately, we can safely assume that the translation is correct for at least one of the WordNet senses. Therefore, we first built Swedish and Bulgarian WordNets that are direct translations of Princeton WordNet by using simple lookups in the dictionaries. This obviously created mistakes but it also made sure that we reuse as much as possible from the existing dictionaries. All entries translated in this way were marked as "guessed".

The next step was to integrate the existing open-source Swedish and Bulgarian WordNets. The problem there is that as we said above, we get good candidates for translations from the synsets, but we still do not know which word matches best a word in another language. We used the following criteria to decide which words from two aligned synsets should be made put into translation relations:

1. In how many senses a pair of words co-occurs in two aligned synsets
2. How many independent dictionaries from PanLex list a pair of words as translations
3. Lower Levenshtein distance between words makes them more appropriate as translations.
4. A pair of derivationally related words in one language should translate to derivationally related words in another language.

Obviously none of the criteria is a silver bullet on its own, but a combination of the four is a pretty good guide. When there is a conflict between the different criteria we give priority to the criterion which is first in the list. Criteria 1 and 3 are also used in Angelov and Lobanov (2016).

The first criteria makes the translation relation more robust for automatic methods. Finding the right translation should be easier even if the precise sense disambiguation is difficult. The second criteria implements the wisdom of the crowd by using PanLex. PanLex is an machine readable accumulation of thousands of dictionaries across many languages. If authors of several dictionaries have included a translation pair then there is a good consensus. The third criteria helps to align together cognates across languages. This is especially helpful for scientific terms which often have Greek or Latin origin.

All of the above criteria can be evaluated numerically and were used to choose how to align the words. However, the automatic alignment is still error prone especially for small data sets. For instance the first criteria is vulnerable if not all of the possible senses of a word are listed in one of the languages. We used the data from WordNet to override the data extracted from the plain GF translation dictionaries. This replaced many of the cases where translations were used in the wrong sense. We also marked those entries as "unchecked", i.e. more certain but validation is still necessary.

Further checking was done by using Wikipedia. The assumption is that for most Wikipedia articles the titles are translations of each other. This means that for every abstract identifier we can look up the article whose titles in different languages match as many of the identifier's translations as possible. After that we checked that the selected article really corresponds to the right word sense. If the WordNet translation in some of the languages does not match the one from Wikipedia then we examined those too. It happens sometimes that the Wikipedia titles are not direct translations of each, so just using the titles to correct translations is not reliable. It can also happen that our lexicon already contains a good translation, while Wikipedia just uses a different synonym for the title.

None of the sources above, nor the automatic alignments guarantee correctness, but this still saves time compared to what we would need if we start from scratch. We incrementally check the correctness of the automatically created entries by considering the same criteria as listed above. The initial entries were created through translation from English, but when checking, we also consider whether the choices for Swedish and Bulgarian are also consistent with each other.

3.3. Morphology

The original WordNet resources are all on the level of lemmas, while in GF we also want complete morphological information for all languages. As we said large morphological lexicons already existed for English and Bulgarian (Angelov, 2014) and for Swedish (Borin et al., 2008). For most words it was enough to look them up in those lexicons. We also used the part of speech as a constraint to avoid the typical noun-verb ambiguities in English.

Whenever we encounter a Swedish word which is not listed in the existing morphological lexicon, then we inflect it by using the smart paradigm in the GF's RGL. For many cases, these paradigms predict the right inflections based on just the part of speech tag and the lemma (Détrez and Ranta, 2012). For Swedish this amounts to 46% of the nouns and

92% of the verbs. The low rate for nouns is due to the unpredictability of the gender in Swedish. While checking the lexicon, we fix the morphological mistakes by providing more forms than just the lemma.

There are no smart paradigms for Bulgarian, so instead, for unknown words we pick the paradigm of a known word with the longest possible common suffix with the target one. Figure 2 shows inflection tables for English, Swedish and Bulgarian generated from the grammar.

3.4. Syntactic Information

In addition to morphology we also need syntactic information. All lexical items, in WordNet are categorized as just nouns, verbs, adjectives or adverbs. The resource grammars, on the other hand, require more fine grained subcategorization that describes how the different words behave syntactically.

The most prominent need of subcategorization is for verbs, since they exhibit different valency patterns. The grammar, for instance, needs to know whether a verb is intransitive, transitive or a two place verb that takes its argument via a preposition. Particle verbs are also treated differently since in translation the particle might be absorbed as a prefix or a suffix of the main verb.

In our lexicon the verbs have different tags depending on their valency. In addition their inflection table includes both the forms of the main verb as well as the eventual particles and prepositions. The current verb subcategorization is based mostly on the uses of the verbs in the corpus of examples coming with WordNet, but there is also an ongoing integration with VerbNet (Schuler, 2005) which will make the verb subcategorization more stable.

Another example is the adverbs which in WordNet are treated uniformly, while in the resource grammars they are categorized according to their syntactic functions. For instance adverbs like "very", which modify adjectival phrases, have different categories than adverbs like e.g. "occasionally", which typically modify verbs. There are also adverbs like "instead" which in one sense can be used on its own while in another it is followed by "of" to form the composite preposition "instead of". In the later case the composite phrase has the category of a preposition in the grammar, while in the original WordNet it is treated as an adverb. The same argument also lets us to treat phrases like "instead of" as multiword expressions, which translate non-compositionally to other languages.

A third apparent example are the numerals, which in WordNet are classified as both nouns and adjectives according to their uses. In the RGL, there is a special of category for numerals, which then can be used syntactically on their own (similar to nouns), in determiners (the cardinals), or like adjectives (the ordinals).

We also use different tags for common nouns and proper names. Nouns in synsets that are linked with the instance relation in WordNet are retagged as proper names in our representation.

3.5. Examples Corpus

In WordNet, most of the glosses in the synsets contain examples in English showing typical uses of the particular senses. We extracted the examples and each example was associated with the word sense that it exemplifies. This corpus is then parsed with GF and sense-annotated.

Note that the Princeton WordNet Gloss Corpus also annotates the examples corpus, but there most words are only labeled with part of speech tags. Sense annotations are available only for the words exemplified with that sentence. In contrast, in order to provide good translations, we sense tagged all words in the corpus. To do that we tagged the example word with the correct sense. All other words we labeled with the first sense for the right part of speech tag. After that when working with the translations we also changed the sense annotations when necessary. For most words the first WordNet sense is also the most typical, and thanks to that in many cases we did not have to change the sense.

An example entry from the corpus is shown on Figure 3. The first line is the GF syntax tree, which is a composition of function applications. Each function is defined separately for each language and describes the rules for building a specific kind of phrases in that particular language. Everything language specific, like word order or gender agreement is thus hidden in the language-specific implementation. The leaves of the tree are the abstract identifiers from the lexicon.

Thanks to the language agnostic representation, from the already constructed GF tree for English, we can generate translations to the other languages. The translation of course will be good only if the lexicon is good. We used that as a way to spot more translation mistakes.

Together with the syntax tree we also store its verbalizations in English, Swedish and Bulgarian. The English sentence is the original example. For Swedish and Bulgarian we first seeded the translations with Google Translate. Later after we validate an entry, we replace the translation with the output from the grammar. The benefit of seeding with automatic translations is that although they may be wrong, when they are correct they serve as an inspiration for which translations fit best the English words.

Storing the both the syntax tree and the verbalizations serves two purposes. First it is a human readable documentation, but second it is also used for regression testing. Whenever we change the lexicon, we also check that the verbalizations still match the trees. In case if they do not, we can eventually reconsider the change.

4. Other Languages

Our work is focused on English, Swedish and Bulgarian, since the development of solid resources for other languages requires language expertise. However, bootstrapping from existing sources without postvalidation can be done automatically. We did that for Catalan, Chinese, Dutch, Estonian, Finnish, Italian, Portuguese, Slovenian, Spanish, Thai and Turkish by using an extension of the alignment method in Angelov and Lobanov (2016) which takes into account the criteria in Section 32. The choice of the languages is determined by the availability of sizeable WordNets, morphological lexicons and resource grammars in GF. For most of the languages we used the aggregated data from Bond and Paik (2012).

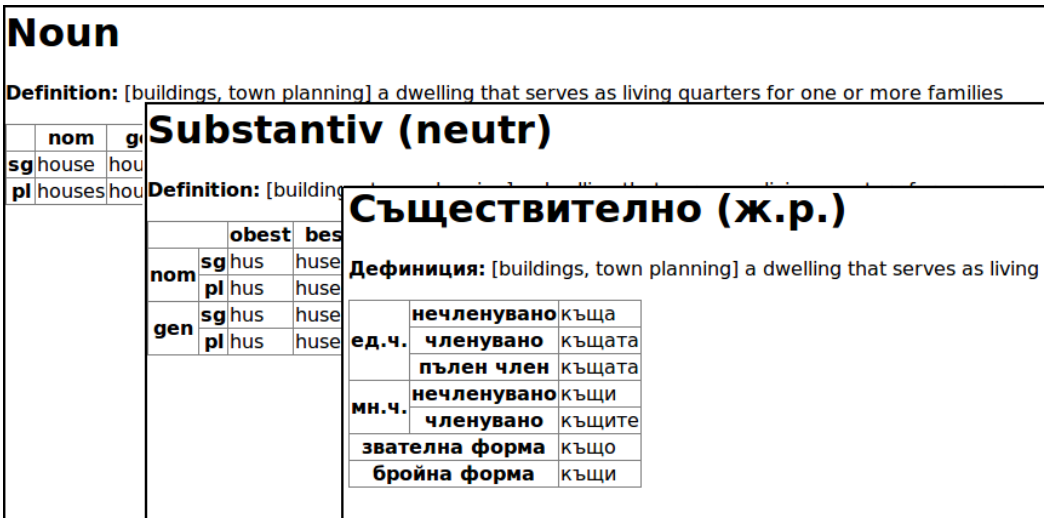

Figure 2: Snapshots of morphological tables generated from GF WordNet

```
abs: PhrUtt NoPConj (UttNP (DetCN (DetQuant IndefArt NumSg)
                                  (AdjCN (PositA rare_1_A)
                                         (UseN word_1_N)))) NoVoc

eng: a rare word
swe: ett sällsynt ord
bul: rjadka duma
key: 1 rare_1_A 00490548-a
```

Figure 3: A parsed and translated example from Princeton WordNet

After the morphology and the WordNet for each language is imported, we also tried to fill in missing translations by using PanLex. This means that we go through each abstract identifier and use its existing verbalizations to find new ones. For each verbalization we construct the list of possible translations in PanLex to all other languages. After that we merge the lists into one by keeping track of how many lists contain a particular translation. The best candidate for each language is selected by using the following criteria:

1. How many of the existing verbalizations have the candidate word as a translation. This is the same as the number of initial lists that contain the candidate.

2. How many independent sources list a pair of words as translations.

3. Lower Levenshtein distance between words makes them more appropriate as translations.

These two steps produce a candidate resource that later must be validated. Some validations, however, can be done automatically with high accuracy. For instance, for abstract identifiers that are linked with Wikipedia we can extract the page titles and check against the translations in lexicon. The Wikipedia titles are generally good translations of each other but as we mentioned in Section 32, there are also exceptions. We did not use that for Swedish and Bulgarian but we used it for the other languages. It gives us a quick way to automatically check translations but there is still small chance for mistakes.

Another way for automatic validation is via PanLex. We know that all words coming from WordNet that are in the same synset share the same sense. What we do not is which one is a translation of which. There are two ways to choose the translation alignment: the criteria in Section 32and the criteria listed above. When the two criteria agree, we can be fairly certain that the choice is right.

All entries satisfying the two requirements above are currently marked as verified, although we are aware that small percentage of errors is possible. Manual validation is required to make everything really trustworthy. The only exception is Finnish, where instead of using automatic alignment and validation, we used the list of translations that is distributed together with the synsets in FinnWordNet. We regard all of these entries as validated, although we know that for Finnish verbs we still lack information for the grammatical cases that must be used with two or three place verbs.

The corpus of examples that is checked for Swedish and Bulgarian has not been checked for any of the other languages.

5. Search Interface

To be able to explore the data we made a web-based search interface[1]. There, it is possible to search for a word in any language and see its senses and synonyms aligned in a matrix as in Figure 1. Clicking on any of the translations in the matrix shows the full inflection table (Figure 2).

[1] https://cloud.grammaticalframework.org/wordnet/

Figure 4: Snapshots of an example from GF WordNet. The highlighting shows the word alignment and at the bottom is shown the gloss for the selected word.

The interface allows searching through the whole lexicon. However, if an entry comes from an existing WordNet, but its translation alignment is not verified yet, then the word is underlined with a yellow line. Similarly, if the word is only generated through a translation dictionary but it is not checked whether it fits the respective sense then it is underlined with red. If there is no special markings then the entry has already been verified.

For each abstract identifier we also show the already verified examples in which the particular sense is used (Figure 4). Since all the sentences are generated from a single abstract syntax tree, it is possible to automatically compute the word and phrase alignments. In the user interface, when the user clicks any of the words in a sentence, the corresponding words (phrases) in the other languages are highlighted as well. Clicking once more, shows alignment one level up in the parse tree. If the current selection covers only a single lexical item, then, below the corpus example, the user interface shows the gloss for that word.

We have also integrated the WordNet Domains resource (Bentivogli et al., 2004) which can be used to search for all words applicable in given domains.

When an identifier is also linked to Wikipedia then we show the thumbnail picture from the article (Figure 5). Clicking on the image leads the user to the actual page.

Showing all the available information in a single page makes it easy to disambiguate and validate new lexical entries. The web interface also allows editing lexical definitions on the web, while the changes are automatically pushed to GitHub[2].

6. Current Status

The status of the WordNet lexicon in February 2020 is shown on Table 1. As can be seen, the only complete language is English but this is because that was our starting point. We just added the necessary adjustments to the word tags and added morphology as explained in Section 33

Most of the work was on Bulgarian and Swedish and the coverage for the verified part of our lexicon is already better than in the original sources, e.g. 8 936 senses in BulTreeBank WordNet and 28 046 senses in the Svenskt OrdNät. For Swedish we have not even managed to integrate all verbs from Svenskt OrdNät. Completing that would increase the coverage even further.

Note that the verified lexicon is still only about half of the full data. There are many correct entries in the rest but we cannot trust them until they are manually checked. Also as we explained in Section 4, for all other languages even the part marked as validated might still contain a small percentage of errors.

Another useful statistics for the lexical coverage is the out of vocabulary words encountered in a large corpus. We evaluated that by using treebanks from the Universal Dependencies project. The statistics are shown in Table 2. When measuring the coverage we ignored all words marked as proper names, numbers, symbols and punctuations. In addition we excluded words that the RGL treates as part of the syntax and not part of the lexicon.

The corpus of examples contains 38,593 sentences of which 11% are currently verified. The verification ensures that the right abstract syntax tree is used, which includes the correct choice for the word senses. We also check that the translation to Swedish and Bulgarian is as good as possible. Very often the translations sound very native, but sometimes when English idioms are used in the source, the translation is too literal. This leads to phrases that are syntactically correct and understandable, but unidiomatic.

We made an experiment where we asked native speakers to post-edit the automatic translations to produce a gold standard for 200 sentences. By comparing the two translations we computed the BLEU scores. It gave us 92.03 points for Bulgarian and 72.69 for Swedish. This shows that although the translation is not perfect, it is still very good.

It is a well known fact that the plain resource grammars in GF are not ideal for direct translation and better results are achieved with specialized application grammars. Still it was encouraging to see that the BLEU scores can be very good.

7. Conclusion

We started building a new WordNet-like resource for Swedish and Bulgarian. Obviously the data is far from complete but it is already of a considerable size (over 30 000 senses). Work is still ongoing and will continue in incremental fashion. The resource is freely available as a library in GF and has already been used in internal projects for rapid application development.

[2] https://github.com/GrammaticalFramework/gf-wordnet

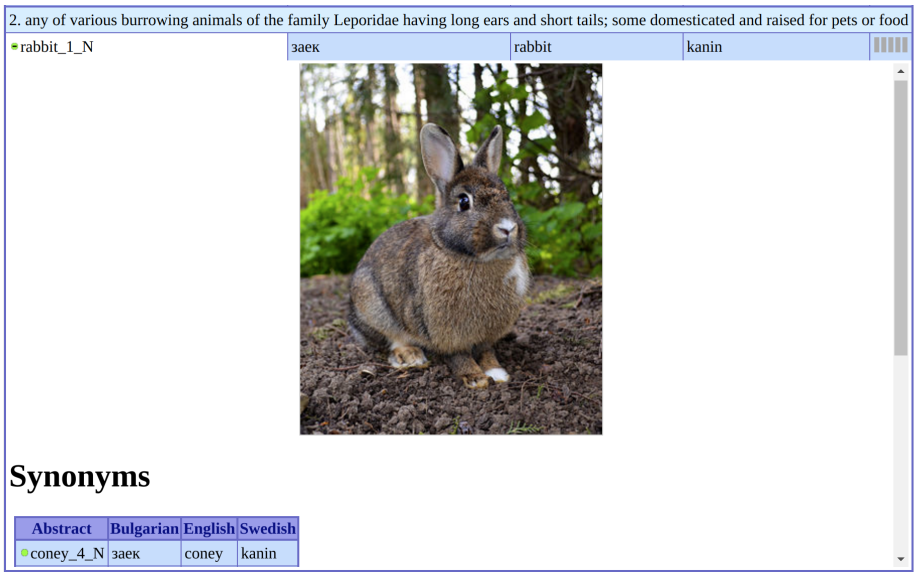

Figure 5: A thumbnail picture from Wikipedia

Language	Senses	Validated	Lemmas
Abstract	107 716	—	—
Bulgarian	**81 377**	**41 411 (51%)**	**33 938**
Catalan	95 663	33 314 (35%)	31 968
Chinese	41 400	1 003 (2%)	27 497
Dutch	95 707	24 818 (26%)	34 761
English	**107 716**	**107 716 (100%)**	**56 648**
Estonian	91 574	14 510 (16%)	36 530
Finnish	97 061	90 677 (93%)	52 166
Italian	96 522	32 141 (33%)	34 479
Portuguese	94 844	55 257 (58%)	33 679
Slovenian	86 424	52 518 (61%)	26 097
Spanish	98 187	35 390 (36%)	32 877
Swedish	**76 898**	**35 820 (47%)**	**33 205**
Thai	64 624	19 347 (30%)	33 573
Turkish	88 896	12 051 (14%)	29 303

Table 1: State of GF WordNet in terms of number of senses, number of validated translations and number of lemmas.

Language	Treebank	OOV%
Bulgarian	**BTB**	**12%**
Catalan	AnCora	18%
Chinese	GSD	75%
Dutch	Alpino	12%
English	**ESL**	**4%**
Estonian	EDT	9%
Finnish	TDT	19%
Italian	ISDT	13%
Portuguese	Bosque	10%
Slovenian	SSJ	18%
Spanish	AnCora	6%
Swedish	**Talbanken**	**5%**
Thai	PUD	24%
Turkish	IMST	31%

Table 2: Percentage of Out of Vocabulary words for different languages/treebanks

We also plan to export the lexicon in the format of the Open Multilingual WordNet project, to make it more easily accessible by projects independent of GF. Similarly for the corpus of examples we have an export to Universal Dependencies (Kolachina and Ranta, 2016).

8. Bibliographical References

Angelov, K. and Ljunglöf, P. (2014). Fast statistical parsing with parallel multiple context-free grammars. In *European Chapter of the Association for Computational Linguistics*, Gothenburg.

Angelov, K. and Lobanov, G. (2016). Predicting translation equivalents in linked wordnets. In *The 26th International Conference on Computational Linguistics (COLING 2016)*, page 26.

Angelov, K. (2014). Bootstrapping open-source English-Bulgarian computational dictionary. In *LREC*, pages 1018–1023.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *In Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources*, pages 101–108.

Bloksma, L., Diez-Orzas, P. L., and Piek, V. (1996). User requirements and functional specification of EuroWordNet project. Deliverable D001, WP1, EuroWordNet, LE2-4003.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Borin, L., Forsberg, M., and Lönngren, L. (2008). The hunting of the BLARK - SALDO, a freely available lexical database for swedish language technology. 01.

Borin, L., Forsberg, M., and Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211, Dec.

Détrez, G. and Ranta, A. (2012). Smart paradigms and the predictability and complexity of inflectional morphol-

ogy. In Walter Daelemans, et al., editors, *EACL*, pages 645–653. The Association for Computer Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Kamholz, D., Pool, J., and Colowick, S. (2014). Panlex: Building a resource for panlingual lexical translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Kann, V. and Hollman, J. (2011). Slutrapport för projektet Folkets engelsk-svenska lexikon. 09.

Koeva, S., G. T. and Genov, A. (2004). Towards Bulgarian wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):45–61.

Kolachina, P. and Ranta, A. (2016). From Abstract Syntax to Universal Dependencies. In *to appear in Linguistic Issues in Language Technology (LiLT)*, volume 13. CSLI, Stanford, August.

Lindén, K. and Carlson, L. (2010). FinnWordNet - WordNet på finska via översättning. *LexicoNordica - Nordic Journal of Lexicography*, 17:119–140.

Ranta, A. (2009). The GF resource grammar library. *Linguistic Issues in Language Technology*.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Department of Computer Science, University of Pennsylvania, December.

Simov, K. and Osenova, P. (2010). Constructing of an ontology-based lexicon for Bulgarian. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Stamou, S., Azer K, O., Pala, K., Christoudoulakis, D., Cristea, D., D Tu, S., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2019). Balkanet: A multilingual semantic network for Balkan languages. 03.

Viberg, o., Lindmark, K., Lindvall, A., and Mellenius, I. (2002). The Swedish WordNet project. In *Proceedings of Euralex*, pages 407–412, August.