Extending the Galician Wordnet Using a Multilingual Bible Through Lexical Alignment and Semantic Annotation

Alberto Simões

Applied Artificial Intelligence Lab (2Ai Lab)
Instituto Politécnico do Cávado e do Ave, Barcelos, Portugal
asimoes@ipca.pt
https://orcid.org/0000-0001-6961-2660

Xavier Gómez Guinovart[1]

Galician Language Technology and Applications (TALG Group)
Universidade de Vigo, Galiza, Spain
xgg@uvigo.gal
https://orcid.org/0000-0001-9961-6953

———— Abstract ————

In this paper we describe the methodology and evaluation of the expansion of Galnet – the Galician wordnet – using a multilingual Bible through lexical alignment and semantic annotation. For this experiment we used the Galician, Portuguese, Spanish, Catalan and English versions of the Bible. They were annotated with part-of-speech and WordNet sense using FreeLing. The resulting synsets were aligned, and new variants for the Galician language were extracted. After manual evaluation the approach presented a 96.8% accuracy.

1 Introduction

WordNet [15, 7] is the key lexical resource in many language applications. While contested by some researchers on the grounds of its fine-grained word senses, it is, indubitably, the most used and replicated lexical knowledge base. For many languages there is, at least, one project trying to build a similar resource.

The main problem is that most of these projects are unable to get enough funding to develop such a large resource by the hand of lexicographers and linguists. This is a bigger problem for under-resourced languages, like Galician. Therefore, these projects employ algorithms to obtain new variants and organise them in synsets, either using complete automatic processes or semi-automatic approaches, where the candidate variants extracted by the algorithms must be manually validated.

Our contribution is another method to obtain candidate variants, applied to the Galician wordnet (Galnet), using other languages wordnets (English, Catalan, Spanish and Portuguese) together with a sentence-aligned multilingual Bible.

To present it, this paper is organised as follows: Section 2 presents the two main resources used for this experiment: Galnet, the Galician wordnet, and the CLUVI multilingual Bible; Section 3 refers to other experiments on creating or enlarging lexical ontologies and their obtained results; in Section 4 the algorithm is explained with some examples; the results obtained by this experiment are then analysed in Section 5, where the different kinds of errors found are detailed; finally, Section 6 draws some conclusions on the methodology used and its applicability to other languages.

2   Language resources

This section presents the two main language resources used in this experiment. First, Galnet, the wordnet for Galician, which is the target for the inclusion of new variants obtained by this experiment. Then, the CLUVI parallel Bible, which will be used as a semantically tagged parallel corpus for the extraction of the new variant candidates.

2.1   WordNet and Galnet

WordNet is a lexical database of the English language, organised as a semantic network where the nodes are concepts represented as sets of synonyms and the links between nodes are semantic relations between lexical concepts. These nodes contain nouns, verbs, adjectives and adverbs grouped by synonymy. In WordNet terminology, a set of synonyms is called a *synset*. The term *variant* was coined during the EuroWordNet project [30] to refer to each (lemmatised) synonym in a synset, which is considered a lexical variant of the same concept. Thus, each synset represents a distinct lexicalised concept and includes all the synonymous variants of this concept. Additionally, each synset may contain a brief definition or gloss, which is common to every variant in the synset, and, in some cases, one or more examples of the use of the variants in context.

In the WordNet model of lexical representation, the synsets are linked by means of lexical-semantic relations. Some of the most frequent relations represented in WordNet are hypernymy/hyponymy and holonymy/meronymy for nouns; antonymy and quasi-synonymy (or semantic similarity) for adjectives; antonymy and derivation for adverbs; and entailment, hypernymy/hyponymy, cause and antonymy for verbs.

Galnet [9] is an open wordnet for Galician, aligned with an interlingual index (ILI) generated from the English WordNet 3.0, following the expand model [30] for the creation of new wordnets, where the variants associated with the Princeton WordNet synsets are translated using different strategies. Galnet can be searched via its own dedicated web interface[2] and can be downloaded in RDF and LMF formats[3].

Galnet is part of the Multilingual Central Repository (MCR)[4], a database that currently integrates wordnets from six different languages (English, Spanish, Catalan, Galician, Basque and Portuguese) using WordNet 3.0 as ILI [12]. Table 1 provides the number of synsets and variants for the different languages gathered in this repository, and their percentage of development with respect to the English WordNet.

▪ **Table 1** Current coverage of wordnets in MCR.

| | English (WordNet 3.0) | | Galician (Galnet 3.0.26) | |
	variants	synsets	variants	synsets
Nouns	146,312	82,115	46,972	31,356
Verbs	25,047	13,767	7,247	3,214
Adjectives	30,002	18,156	10,423	6,375
Adverbs	5,580	3,621	1,651	1,079
Total	206,941	117,659	66,293	42,024
%	100%	100%	32,0%	35%
	Spanish (MCR 2016)		**Portuguese (MCR 2016)**	
	variants	synsets	variants	synsets
Nouns	101.027	55.227	17.125	10.047
Verbs	20.953	9.541	8.360	3.786
Adjectives	20.938	12.373	6.330	3.581
Adverbs	3.583	1.854	789	528
Total	146.501	78.995	32.604	17.942
%	70,8%	67,1%	15,8%	15,2%
	Catalan (MCR 2016)		**Basque (MCR 2016)**	
	variants	synsets	variants	synsets
Nouns	73,810	46,917	40,420	26,710
Verbs	14,619	6,349	9,469	3,442
Adjectives	11,212	6,818	148	111
Adverbs	1,152	872	0	0
Total	100,793	60,956	50,037	30,263
%	48,7%	51,8%	24,2%	25,7%

2.2 The CLUVI multilingual Bible

The CLUVI Corpus[5] is an open collection of human-annotated sentence-level aligned parallel corpora, originally designed to cover specific areas of the contemporary Galician language in relation to other languages. With over 47 million words, the CLUVI collection currently comprises twenty-one parallel corpora in nine specialised registers or domains (fiction, computing, popular science, biblical texts, law, consumer information, economy, tourism, and film subtitling) and different language combinations with Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque, German, Latin and Chinese.

The CLUVI search application allows for very complex searches of isolated words or sequences of words, and shows the bilingual equivalences of the terms in context, as they appear in real and referenced translations. When the term search is a lemma in a language, the result texts could include WordNet-based suggestions on its lexical equivalences in the translation languages using colour codes. In addition, the legal section of the CLUVI corpus, a subset of Galician-Spanish legal texts with 6.5 million words, can be freely downloaded with CC BY-NC-SA 3.0 license[6].

---

[5] http://sli.uvigo.gal/CLUVI/
[6] http://hdl.handle.net/10230/20051

At the moment, the CLUVI is the parallel corpus that contains the largest number and the most varied thematic range of translations from/to the Galician language. Galician texts present in the CLUVI collection sum up to about 12,000,000 words, which means a quarter of the total of the tokens in the corpus for all the languages and domains of translation. By way of comparison, other parallel corpora that currently facilitate access to translations to the Galician language are the collection of downloadable corpora from OPUS Project[7] [29] and from the Per-Fide Corpus[8] [3]. On the one hand, the OPUS collection provides Galician translated texts, mainly from English, taken from the web and automatically aligned at sentence level. Galician texts in OPUS total around 7,600,000 tokens, coming mainly from the localisation of the Linux operating system environment (Gnome, KDE4 and Ubuntu). On the other hand, the Per-Fide collection includes Portuguese-Galician software localisation parallel texts derived from the OPUS Project with about 400,000 tokens in Galician.

The multilingual Bible built for the CLUVI Corpus aligns the translations of the biblical texts in thirteen linguistic variants – Latin, Galician, Brazilian Portuguese, European Portuguese, Catalan, French, Italian, Spanish, English, German, Basque, Simplified Chinese and Traditional Chinese – , with a total of 31,279 translation units, generally equivalent to the Bible verses, and 7,481,611 words: 535,423 words in Latin, 656,998 words in Galician, 770,410 words in Brazilian Portuguese, 662,853 words in European Portuguese, 723,194 words in Catalan, 719,229 words in French, 674,795 words in Italian, 706,125 words in Spanish, 759,824 words in English, 701,279 words in German, 505,043 words in Basque, 31,308 words in Simplified Chinese and 35,130 words in Traditional Chinese. The CLUVI multilingual Bible collects the translations of all the books shared by the Western biblical canon (Protestant, Lutheran, Anglican and Roman Catholic traditions), including the Old Testament (Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Judges, Ruth, 1 and 2 Samuel, 1 and 2 Kings, 1 and 2 Chronicles, Ezra, Nehemiah, Esther, Job, Psalms, Proverbs, Ecclesiastes, Song of Songs, Isaiah, Jeremiah, Lamentations, Ezekiel, Daniel, Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah and Malachi) and the New Testament (Matthew, Mark, Luke, John, Acts, Romans, 1 and 2 Corinthians, Galatians, Ephesians, Philippians, Colossians, 1 and 2 Thessalonians, 1 and 2 Timothy, Titus, Philemon, Hebrews, James, 1 and 2 Peter, 1, 2 and 3 John, Jude and Revelation).

There have been some other projects focused on the creation of an aligned multilingual Bible corpus for linguistic research, such as [24] and [6]. However, up to now, the CLUVI multilingual Bible represents the only available parallel version of the Scriptures which includes a Galician translation.

3   Related work

The *expand model* mentioned above and followed by Galnet for the creation of the Galician wordnet has also been used in the development of the wordnets for Italian [18], Indonesian [21], Hungarian [14], Croatian [22] French – WOLF [25] and WoNeF [20] wordnets – and Kurdish [2]. The same approach has been taken in the MCR framework for the creation of the wordnets of Spanish [4], Catalan [5], Basque [19] and Portuguese [26].

In the expand model, one of the main methodologies used to extend a wordnet coverage from the variants associated with the Princeton WordNet synsets is the acquisition of their translations from parallel corpora. In fact, we have applied that methodology in a previous

---

[7] `http://opus.lingfil.uu.se/`
[8] `http://per-fide.di.uminho.pt/`

■ **Table 2** Precision of parallel corpus-based expansion in [8] and number of obtained variants.

	Precision	New variants
SemCor	78.13%	2,053
Unesco	80.84%	2,150
Lega	77.42%	1,172
Eroski	80.28%	1,777
Tectra	82.74%	948

phase of the Galnet development [8], using the WN-Toolkit [16] – a set of Python programs for the creation or enlargement of wordnets – to expand the Galnet first distribution (released for download in 2012 as part of the MCR 3.0) from two different available parallel textual resources: the automatically translated (with Google Translate) English–Galician SemCor Corpus[9]; and the four sections of the CLUVI Corpus, namely, the Unesco Corpus of Spanish–Galician scientific-technical texts, the Lega Corpus of Galician–Spanish legal texts, the Eroski Corpus of Spanish–Galician consumer information texts and the Tectra Corpus of English–Galician literary texts. In all cases, only the English or Spanish part of the parallel corpora has been sense-tagged for the experiment, using FreeLing [17] for parsing with the UKB word sense disambiguator [1].

After the sense annotation, a word alignment algorithm is used in order to identify the candidate target variants in the texts. For that experiment, we use a very simple word alignment algorithm which is based on the most frequent translation and which is available as a part of the WN-Toolkit. That alignment algorithm calculates the most frequent translation found in the corpus for each synset taking into account that the parallel corpus must be tagged with WordNet synsets in the source part and the target corpus must be lemmatised and tagged with very simple tags (*n* for nouns, *v* for verbs, *a* for adjectives, *r* for adverbs, and any other letter for other parts of speech).

In Table 2 we can observe the precision and the number of new variants obtained in the experiment from each parallel corpus. The evaluation has been performed in an automatic way, comparing the obtained variants with the existing variants in the current distribution of Galnet. If the variant obtained for a given synset already exists in that same synset, the result is evaluated as correct. If there are no Galician variants for a given synset in the reference Galnet, this result is evaluated as incorrect. It should be noted that the automatically obtained precision values tend to be lower than real values. The reason is that sometimes we have one or more variants for a given synset in the reference Galnet, but the obtained variant is not present. If the obtained variant turns out to be correct, it will be evaluated as incorrect anyway.

In [16] the same methodology is applied to the automatic translation (via Google Translate) of the English SemCor to six languages (Catalan, Spanish, French, German, Italian and Portuguese). Table 3 shows the results of the automatic evaluation of the data yielded from that experiment for the expansion of the wordnets of these six languages.

Another line of research on automatic extension of ontologies is carried out for Portuguese in the framework of Onto.PT[10] [11], where new synsets are obtained from lexicographical, encyclopedic and textual resources using different similarity indexes for lexical acquisition [10].

---

[9] `http://www.gabormelli.com/RKB/SemCor_Corpus/`
[10] `http://ontopt.dei.uc.pt/`

■ **Table 3** Precision of parallel SemCor-based expansion in [16] and number of obtained variants.

	Precision	New variants
Catalan	87.63%	449
Spanish	88.93%	504
French	91.83%	142
German	70.26%	1,285
Italian	93.81%	66
Portuguese	84.14%	324

■ **Listing 1** Portuguese segment after being processed by FreeLing.

```
Javé javé NCMS000 1 −
concedeu conceder VMIS3S0 1 02327200−v:0.00125866/02199590−v:0.0012442
uma um DI0FS0 0.903495 −
grande grande AQ0CS00 0.998339 01382086−a:0.00262735/01472628−a:0.000829225
vitória vitória NCFS000 1 07473441−n:0.0155845
```

4 Methodology

The process of extraction of proposals for variants uses, as resources, the multilingual Bible, namely the alignments from Portuguese, English, Catalan and Spanish to Galician, as well as these languages' respective wordnets. Section 4.1 describes how the Bible was preprocessed, and Section 4.2 explains the variant proposals extraction algorithm.

4.1 Multilingual Bible Preprocessing

The multilingual Bible is available in a set of translation memory exchange (TMX) files, one for each Bible book. Thus, the first step is the concatenation of these files, in a single translation memory, and the projection of each language in a separate textual file, where each translation unit resides in its own line. Also, as a few translation units do not include translations for every other language, in these cases an empty line is created in the respective text file.

Each one of these files is then processed using FreeLing, where each word is annotated with its lemma and part of speech. For those words present in the wordnet for that language, FreeLing uses the UKB algorithm in order to associate the more probable sense (which corresponds to a WordNet synset) to that word. FreeLing does not add only one sense, but a set of different senses, adding a probability measure to each one. This process is done taking care of translation units boundaries, in order to keep them after the annotation process. Listing 1[11] presents an example of a Portuguese translation unit after being processed by FreeLing.

In Listing 1, each line corresponds to a word, with different fields (or columns) separated by spaces: original word, lemma, part-of-speech, tagger confidence and a list of references to the synsets containing that lemma. References are separated by a slash, and each reference comprises the synset ILI and the UKB algorithm confidence.

---

[11] The example was truncated to a maximum of two senses.

Listing 2 Portuguese segment based on lemmas.

```
javé conceder um grande vitória
```

$$vitória \begin{cases} vitoria & : & 56.37 \ \% \\ salvación & : & 7.78 \ \% \\ triunfo & : & 4.66 \ \% \\ xustiza & : & 2.56 \ \% \\ axuda & : & 1.77 \ \% \end{cases}$$

Figure 1 Example of PTD entry for the word *vitória* for the language pair PT ↔ GL.

The lemma information from these files are then concatenated together, in order to reconstruct the original text files, but with each word replaced by each lemma (Listing 2).

These lemmatised files are then used by NATools [27] to compute probabilistic translation dictionaries (PTD) for the following language pairs: PT ↔ GL, EN ↔ GL, CA ↔ GL and ES ↔ GL. A PTD is a dictionary that maps each word from the source language to a set of probable translations, together with their probability (see Figure 1 for an example entry). The main advantage of using this kind of dictionary is that the domain and range of the dictionary cover all the corpus words (even if, in some situations, proposing a bad translation).

4.2 Extraction of Variant Proposals

The extraction of variant proposals is done through the alignment of words in the Bible that were annotated by FreeLing with one or more WordNet senses. The algorithm is based on the following steps, which are executed for each translation unit of the Bible:

1. For each one of the source languages (PT, EN, CA and ES) the algorithm gets the current annotated translation unit, and finds semantically annotated words. For each one of these words, it saves in a dictionary the ILI for the three most probable senses. This step results in a dictionary that can be formally defined as the following map:

   $$\text{ILI} \mapsto \left(\mathcal{L} \mapsto \langle \text{W}_{\mathcal{L}} \cup \text{L}_{\mathcal{L}} \mapsto \mathbb{N} \rangle \right),$$

   where $\mathcal{L}$ stands for the set of source languages, and $\text{W}_{\mathcal{L}}$ and $\text{L}_{\mathcal{L}}$ for the set of words and lemmas, respectively, from language $\mathcal{L}$. These forms and lemmas are counted and their number of occurrences is saved.

   As an example, for the synset with ILI = 07473441-n, the top level dictionary would have the following entry, meaning that only one translation was found in each language and that this synset occurs two times in the whole Bible:

   $$07473441\text{-n} \rightarrow \begin{cases} \text{PT} & \rightarrow & \{\text{vitória} \rightarrow 2 \\ \text{EN} & \rightarrow & \{\text{victory} \rightarrow 2 \\ \text{ES} & \rightarrow & \{\text{victoria} \rightarrow 2 \\ \text{CA} & \rightarrow & \{\text{victòria} \rightarrow 2 \end{cases}$$

2. The dictionary created in the previous step is processed, one entry at a time (for each ILI): if that ILI has at least three source languages – that is, if the previous process found this ILI in, at least, three of the four processed languages – then it is considered. If not, it is discarded.

For each one of these forms and lemmas, the probabilistic translation dictionary is queried, and the probable translations retrieved. Those translations with a translation probability lower than 20% are discarded. For the other translations, another dictionary is created that maps each possible translation to the accumulated translation probability and an accumulated similarity measure (a simple Boolean measure that states if the source word and translation word are orthographically similar). This dictionary has the following structure:

$$\text{W}_{GL} \mapsto \langle \mathbb{R}, \mathbb{N} \rangle$$

where the first element of the pair is the accumulated translation probability, and the second is the accumulated similarity measure.

For the ILI presented before, this dictionary would contain:

$$\begin{cases} \text{triunfo} & \rightarrow & (0.2024, 0) \\ \text{vitoria} & \rightarrow & (2.4876, 3) \end{cases}$$

which means the Galician variant *vitoria* is similar to three variants from any other languages, and has a greater accumulated translation probability than the word *triunfo*. Finally, the target sentence, from the translation unit, is retrieved. Each word from the sentence whose part of speech matches the part of speech from the ILI (note that ILI includes a character, at the end of the code, indicating the PoS for those synset variants) and also matches one of the previously obtained translations will be considered a variant. If this variant is already part of the Galnet synset, then it is marked as a known variant (which will be used to validate the algorithm). All these variants are saved in a global dictionary, as variant candidates for that specific ILI. This dictionary also saves an accumulated similarity measure, and a counter on how many translation units suggested this specific candidate:

$$\text{ILI} \mapsto (\text{L}_{GL} \mapsto \langle \mathbb{N}, \mathbb{N} \rangle)$$

For the example ILI used above, the resulting dictionary includes:

$$07473441\text{-n} \rightarrow \begin{cases} \text{triunfo} & \rightarrow & (2, 0) \\ \text{vitoria} & \rightarrow & (14, 37) \end{cases}$$

meaning that the word *triunfo* was suggested performing the algorithm in two segments of the Bible, while the word *vitoria* was suggested by fourteen segments. Also, *triunfo* was never similar to any other language variant while *vitoria* was similar 37 times (see Section 4.2.1).

4.2.1 Similarity Measure

Regarding the similarity measure mentioned before, its main goal is to give preference to those variants that are orthographically similar to variants from other languages. This is specially useful given that three from the four source languages (PT, CA and ES) have some resemblances. For the pair PT $\leftrightarrow$ EN, the dictionary computed on some previous experiments for the Portuguese wordnet [28], based on rewriting rules, is used. For the other languages, the well known Levenshtein distance [13] is applied, with a distance of 1.

5   Evaluation and results

For the evaluation of this methodology we designed a protocol oriented to the manual analysis of the results by an expert lexicographer. The review of the candidates is done by checking their lexicographical adequacy to the concept represented in the WordNet knowledge base. In case of lack of adequacy, the proposal receives a code that indicates the reason for its exclusion. The codes established in the revision protocol are the following:

1. **MCR source:**
   wrong variant introduced by mistakes in the wordnets for other languages;
2. **Lemmatisation:**
   wrong variant introduced by a mistake in the lemmatisation process;
3. **Galician normative:**
   variant whose form is deprecated by the official Galician normative;
4. **False friends**:
   variant form is equal to a variant in other language, but with a different meaning;
5. **Levenshtein distance:**
   similar to the false friend class, but originated by the similarity measure.

5.1   Error analysis

In the following sections, we will describe and exemplify the phenomena grouped by each error code, and their influence on the overall evaluation of the results.

5.1.1   MCR source

The existing lexicographic errors found in the MCR – especially in Portuguese, Catalan or Spanish – can produce erroneous candidates for extraction. These candidates are not considered errors of the extraction methodology and, therefore, are not taken into account for the evaluation of its results.

For example, the Portuguese variant "fazer" included in the concept "give birth" (ili-30-00056930-v)[12] is too generic for this sense and gives rise to the proposal "facer", also too generic in Galician and therefore classified as incorrect.

The identification of lexicographic errors in Portuguese through this protocol will also be used for the revision and maintenance of the PULO knowledge base.

5.1.2   Lemmatisation

In some cases, the wrong candidates come from the incorrect lemmatisation performed by FreeLing. These cases can not be considered as errors of the extraction methodology, so they will not be included in the evaluation of its performance.

By way of illustration, some of these cases are the wrong candidates "bendiciu" (past tense inflected form of the verb) instead of "bendicir" (infinitive of the verb by which it must be lemmatised); "costa" instead of "costas" for the concept in ili-30-05588174-n,[13] where the lemma should go in plural because the word can only be used in the plural form in this

---

[12] http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-00056930-v
[13] http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-05588174-n

sense; or "testemuño" instead of "testemuña" for the concept in ili-30-10786517-n,[14] where the lemma should be feminine because the word can only be used in feminine in this sense.

In all these cases, the extraction algorithm generates a wrong candidate for Galnet because of the bad selection of the lemma during the automatic linguistic analysis carried out by FreeLing.

5.1.3   Galician normative

In a few cases, the Galician candidate generated from the CLUVI multilingual Bible represents a variant rejected by the current official normative of the Galician language [23]. For example, the proposed candidate "lá", an ancient spelling of the word for the concept of "outer coat of especially sheep and yaks" (ili-30-01899593-n),[15] is not well written following the current regulations, which prescribe it without graphic accent (i.e., "la").[16]

These erroneous candidates cannot be considered as the result of any dysfunction of the extraction methodology, so they cannot be taken into account for the evaluation of the accuracy of their results.

5.1.4   False friends

In some cases, an erroneous candidate is produced because of the orthographic identity – but not identity of meaning – between the proposed Galician form and its source. These false friends occur more frequently between Galician and Portuguese, although they can also occur between Galician and Catalan or Spanish. For example, the erroneous proposal "apagar" for the meaning "put an end to; kill" (ili-30-00478217-v)[17] is generated from the formal identity of the Galician word with the Portuguese form "apagar" which, unlike Galician, does have that meaning.

The cause of this erroneous behaviour lies in certain errors of the semantic tagging of the UKB algorithm and it is limited to polysemous words that are false friends in some of their meanings, but are true friends in others. For example, the word "apagar" in Galician is polysemous. In the example above (ili-30-00478217-v), it is a false friend of the Portuguese verb "apagar". However, in another of its senses, "put out, as of fires, flames, or lights" (ili-30-02761897-v), the Portuguese verb "apagar" and the Galician verb "apagar" are true friends, because they share the same meaning. In the parallel corpus, a number of cases of "apagar" are classified with the shared sense of ili-30-00478217-v in equivalent phrases from Galician and Portuguese. But it also happens that the Galician word "apagar" with the ili-30-02761897-v sense is incorrectly classified by FreeLing as ili-30-00478217-v.

Because of this erroneous semantic tagging, attributable to a malfunction of the UKB algorithm, the extraction algorithm mistakenly proposes the candidate "apagar" as a candidate Galician translation for the sense of ili-30-00478217-v. Hence, these wrong candidates can not be considered as extraction errors, as they are due to the UKB algorithms, and as such, they will not be taken into account in the evaluation of the results of the proposed algorithm.

---

[14] http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-10786517-n

[15] http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-01899593-n

[16] https://academia.gal/dicionario/-/termo/busca/la

[17] http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-00478217-v

Table 4 Results of the human evaluation.

Candidatures	1665
No error	1443
MCR source	144
Lemmatisation	8
Galician normative	4
False friends	18
Levenshtein distance	48
Precision	96.8%

5.1.5   Levenshtein distance

Most of the inaccuracies of the methodology presented here are found in candidates formally similar to the source variants, but with different meaning. For example, the system generates the Galician proposal "ver" for the concept "come to pass; arrive, as in due course" (ili-30-00341917-v)[18] due to its formal similarity with the Portuguese variant "vir" for this sense, with which it maintains a Levenshtein distance 1, accepted by the system in words between 1 and 5 letters. However, "ver" ("to see", in Galician) does not have that sense at all, so it is an unexpected error generated by the methodology. The correction of these errors in a programmatic way is not easy, since the elimination of Levenshtein distance would suppose an unwanted decrease in the coverage of the results.

5.2   Evaluation results

After running the extraction algorithm we obtained 4,353 new Galician variants, from which 1,665 were orthographically similar to variants from other languages according to the similarity measure referred to in section 4.2. As previously stated, the evaluation of the results has been carried out completely manually by an expert lexicographer by checking the lexicographical adequacy of the candidates to the concept represented in the WordNet knowledge base, and in case of lack of adequacy, indicating with the corresponding error code the reason for its exclusion. In Table 4 we can observe the results of this evaluation.

As said before, precision is calculated without taking into account the existing lexicographic errors found in the MCR, the errors from the lemmatisation process performed by FreeLing, the errors from changes in the official normative of Galician or the errors in semantic tagging produced by the UKB algorithm, which are not considered inherent errors of the extraction methodology.

6   Final remarks

The results of human evaluation in section 5 show that the presented methodology outperforms the results reported in previous works [8, 16] and discussed above in Section 3 (Tables 2 and 3). This would demonstrate the importance of associating lexical semantic information to lexical alignment for the identification of variant candidates in parallel corpora.

---

[18] `http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-00341917-v`

Future work includes both introducing the validated set of Galician extracted variants in the Galnet knowledge base, as generating the Galnet 3.0.27 new distribution.[19] We will also revise the erroneous Portuguese variants in PULO identified during the evaluation. Finally, we will apply this methodology for the expansion of data in the PULO wordnet.

───── **References** ─────

1   Eneko Agirre and Aitor Soroa. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 33–41, 2009.

2   Purya Aliabadi, Mohamed Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. Towards building KurdNet, the Kurdish WordNet. In *Proceedings of the 7th Global Word-NetConference*, Tartu, Estonia, 2014.

3   José João Almeida, Sílvia Araújo, Nuno Carvalho, Idalete Dias, Ana Oliveira, André Santos, and Alberto Simões. The Per-Fide Corpus: A new resource for corpus-based terminology, contrastive linguistics and translation studies. In Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira, editors, *Working with Portuguese Corpora*, pages 177–200, London, 2014. Bloomsbury Publishing.

4   Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. Combining multiple methods for the automatic construction of multilingual Word-Nets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338, 1997.

5   Laura Benítez, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. Methods and tools for building the Catalan WordNet. In *In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages*, 1998.

6   Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395, 2015.

7   Christiane Fellbaum, editor. *WordNet: An electronic lexical database.* MIT Press, Cambridge, 1998.

8   Xavier Gómez Guinovart and Antoni Oliver. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50, 2014.

9   Xavier Gómez Guinovart and Miguel Anxo Solla Portela. Building the Galician wordnet: methods and applications. *Language Resources and Evaluation*, 52, 2017. `doi:10.1007/s10579-017-9408-5`.

10   Hugo Gonçalo Oliveira. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese.* Tese de doutoramento, Universidade de Coimbra, 2013. URL: `http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_PhdThesis2012.pdf`.

11   Hugo Gonçalo Oliveira and Paulo Gomes. ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393, 2014. `doi:10.1007/s10579-013-9249-9`.

12   Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual Central Repository version 3.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, 2012. ELRA.

13   Vladimir Levenshtein. Binary Codes Capable of Correcting Deletions and Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

───────────────

[19] `http://sli.uvigo.gal/download/SLI_Galnet/`

14 Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. Methods and results of the Hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference. GWC*, pages 387–405, Szeged, Hungary, 2008.

15 George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

16 Antoni Oliver. WN-Toolkit: Automatic generation of wordnets following the expand model. In *Proceedings of the 7th Global WordNet Conference*, Tartu, 2014. GWN.

17 Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

18 Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet. developing an aligned multilingual database. In *1st International WordNet Conference*, pages 293–302, Mysore, India, 2002.

19 Elisabete Pociello, Eneko Agirre, and Izaskun Aldezaba. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142, 2011. `doi: 10.1007/s10579-010-9131-y`.

20 Quentin Pradet, Gaël de Chalendar, and Jaume Baguenier Desormeaux. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia, 2014.

21 Desmond Darma Putra, Abdul Arfan, and Ruli Manurung. Building an Indonesian Word-Net. In *Proceedings of the 2nd International MALINDO Workshop*, 2008.

22 Ida Raffaeli, Bekavac Božo, Željko Agić, and Marko Tadić. Building Croatian WordNet. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary, 2014.

23 Real Academia Galega. *Normas ortográficas e morfolóxicas do idioma galego*. Editorial Galaxia, Vigo, 2004.

24 Philip Resnik, Mari Broman Olsen, and Mona Diab. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2):129–153, 1999.

25 Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex*, 2008.

26 Alberto Simões and Xavier Gómez Guinovart. Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *Lecture Notes in Computer Science*, pages 239–248, Berlin, 2014. Springer.

27 Alberto Simões and José João Almeida. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September 2003.

28 Alberto Simões and Xavier Gómez Guinovart. Dictionary Alignment by Rewrite-based Entry Translation. In José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors, *2nd Symposium on Languages, Applications and Technologies*, volume 29 of *OpenAccess Series in Informatics (OASIcs)*, pages 237–247, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. `doi:10.4230/OASIcs.SLATE.2013.237`.

29 Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214—-2218, Istanbul, 2012. ELRA.

30 Piek Vossen, editor. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, 1998.