Building a Spanish-German Dictionary for Hybrid MT

Anne Göhring
Institute of Computational Linguistics
University of Zurich
goehring@cl.uzh.ch

Abstract

This paper describes the development of the Spanish-German dictionary used in our hybrid MT system. The compilation process relies entirely on open source tools and freely available language resources. Our bilingual dictionary of around 33,700 entries may thus be used, distributed and further enhanced as convenient.

1 Introduction

Nowadays it is possible to set up a baseline SMT system for any language pair within a day, given enough parallel data, as well as the software to train and decode, is freely available. Whereas SMT systems profit from large amounts of data, following the general motto "more data is better data", the rule-based MT systems on the other hand benefit from high quality data. Developing a hybrid MT system on a rule-based architecture[1], one of our aims is to build and extend a high quality Spanish-German dictionary. We focus on the unidirectional lexical transfer from Spanish to German, as we are translating only in this direction. We want to balance the disadvantage of rule-based systems with respect to lexical coverage when compared to statistical MT systems trained on large scale corpora. To achieve this goal, we have merged existing resources into one bilingual dictionary. As a result we now have a consolidated Spanish-German dictionary of around 33,700 entries.

In the following section, we will give an overview of resources for German and Spanish related to our work. In section 3 we will explain which resources we used and how we combined them. We will also present some figures about the coverage of the resulting bilingual dictionary. Section 4 is dedicated to specific German linguistic issues we have addressed to complete our dictionary with the necessary morphological information. In the last section, we present our ideas for future work.

[1]Our system is derived from Apertium/Matxin, and so is the dictionary format (see 3.1).

2 Related work and resources

Many monolingual and bilingual resources for Spanish and German already exist, some are publicly available, others only under license. The web services Canoo, Leo and Systran are freely accessible but prohibit any automated content extraction. Also the German wordnet GermaNet restricts its usage to the academic community. The HyghTra project develops hybrid high quality translation systems based on commercial resources provided by Lingenio, a language tool company specialized in machine translation (Babych et al., 2012).

In our project we work on similar systems but we follow a free resources and open source policy. This is the case of the open source suite of language analyzers FreeLing (Padró and Stanilovsky, 2012), which offers a Spanish dictionary that contains over 550,000 full-fledged word forms. The bilingual dictionary "ding-es-de"[2] compiled for the "ding" dictionary lookup program provides more than 21,000 entries.

Besides lexicons, other types of resources may provide us with extra material. Escartín (2012) has built a Spanish-German corpus with the specific aim to study multiword expressions in a translation context. There are larger parallel corpora like Acquis JRC, Europarl (Koehn, 2005), and MultiUN (Eisele and Chen, 2010), and also different multilingual wordnets such as BabelNet (Navigli and Ponzetto, 2012) and the Multilingual Central Repository (Gonzalez-Agirre et al., 2012).

[2]savannah.nongnu.org/projects/ding-es-de

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 30–35,
Gothenburg, Sweden, April 27, 2014.

Yet another kind of valuable resources are the monolingual and parallel treebanks like the Spanish AnCora (Taulé et al., 2008) and IULA treebanks (Marimon et al., 2007), the German TiGer (Brants et al., 2004), the multilingual 'universal dependency treebank' (McDonald et al., 2013), and the Spanish-German SQUOIA treebank (Rios and Göhring, 2012).

All the open resources listed above have played or will play a role in building and extending our bilingual dictionary.

3 Compilation of a Spanish-German dictionary

3.1 Format

As we started our machine translation project using the Apertium/Matxin framework (Mayor et al., 2012), we adopted its dictionary format. Though the XML format is specific to our application, it is per definition easy to adapt. As shown in Fig. 1, a bilingual entry `<e>` has at least a left and a right side, `<l>` and `<r>` respectively, and this pair typically refers to a paradigm `<par>`. Furthermore, attributes can be set to whole paradigms as well as to individual entries. We have defined general and more refined paradigms to represent the German morphological classes and the features we need for generating the correct word forms.[3]

```
<e><p>
      <l>nota</l>
      <r>Bemerkung</r>
   </p><par n='NC_NN_FEM'/>
</e>
<e><p><l>nota</l>
      <r>Hinweis</r>
   </p><par n='NC_NN_MASC'/></e>
```

Figure 1: Two entries of the Spanish common noun *nota* (en: note; grade, mark).

3.2 Synonyms and polysemous words

Often a Spanish word has many German translations, and vice versa. This fact is of course reflected in our dictionary, where a Spanish lexical unit (a lemma of a given part-of-speech) has multiple entries, i.e. different corresponding German lexical units.

Fig. 2 shows the same example as in Fig. 1, the polysemous Spanish noun *nota*, together with German translations grouped according to the different senses. Note that the German word *Note* is not always the correct translation as it does not entail all senses: it is not a valid translation for sense 1.

nota ⇒
- Bemerkung, Hinweis, Notiz (sense 1: memo, note, notice)
- *Note*, Schulnote, Zensur (sense 2: mark, grade)
- Musiknote, *Note* (sense 3: musical note)

Figure 2: Different senses of the Spanish noun *nota* and their corresponding German translations.

On the one hand, the dictionary should contain as many word translations as possible in order to achieve a high coverage for both languages. On the other hand, the more fine-grained the choices in the lexicon are, the harder the lexical disambiguation becomes (Vintar et al., 2012). Although hand-written selection rules narrow down the choice in specific cases, machine learning approaches are required in order to make better lexical choices in general.

3.3 First compilation

We first merged the entries of the "ding-es-de" dictionary to the translations of the AnCora/FreeLing vocabulary we obtained by crawling the Spanish Wiktionary in 2011. Since this first compilation period, we have manually added new entries as required by the development of our MT system. At the end of 2013, the collected bilingual entries for the open classes noun, verb, adverb and adjective amounted to 25,904 (see Tab. 1).

At this point we decided to systematically extend our bilingual dictionary and evaluate its coverage . Translating from Spanish to German, we are first of all interested in the coverage of the source language Spanish. Compared to the more than 88,000 lemmas with about double as much senses contained in the DRAE[4], our bilingual dictionary covers not even 5% of the monolingual entries. But the DRAE is a reference dictionary, with certain shortcomings such as missing the newest neologisms and keeping obsolete words in its lexicon. Furthermore, it is not a free resource.

[3] See also Fig. 4 in section 4.2.

[4] Diccionario de la Real Academia Española; 22nd edition DRAE (2001); see `www.rae.es`.

3.4 Exploiting Wiktionary and BabelNet

FreeLing's Spanish lexicon contains 49,477 lemmas of common nouns and 7649 verb lemmas. Before the addition of more data, our dictionary covered only 19.44% of FreeLing's nouns and 22.9% of its verbs. Crawling the Wiktionary pages for the missing lemmas, we collected no more than 309 additional noun and 78 verb entries. Due to this marginal increase, we decided to test other sources. Through BabelNet's API we were able to extract 21,587 German translations of 13,824 Spanish common nouns. We used the morphology tool mOLIFde (Clematide, 2008) to analyze the German side of these BabelNet word pairs. We discarded those pairs that did not receive any analysis. The remaining candidate entries amount to 7149. Though we have not yet assess the quality of this material, the observed coverage gain from these potential bilingual entries looks promising. Adding entries for 5528 Spanish nominal lemmas increases the coverage of common nouns by more than 11% (see Tab. 1).

es-de.dix	**end 2013**	**+ new**	**current**
	Spanish-German entries		
noun	16,136	7,149	23,285
verb	4,256		4,256
adverb	316		316
adjective	5,196	640	5,836
Total	25,904		**33,693**
	Unique Spanish lemmas		
noun	10,559	5,528	16,087
adjective	3,029	627	3,656

Table 1: Size of the Spanish-German dictionary at the end of 2013 and after adding entries extracted from BabelNet.

Starting with the vocabulary extracted from a corpus of European Spanish newspaper texts, we expect our bilingual dictionary to be biased with respect to the language variety, register and genre. In our MT project we focus on Peruvian Spanish. Therefore, we want to measure the specific lexical coverage for this variety. In a first step, we compared our Spanish-Quechua dictionary with the Spanish-German lexicon by computing the overlap of their Spanish vocabularies. Only 50% of the 2215 single Spanish verbs with a Quechua translation also have a German equivalent. Crawling Wiktionary for the untranslated 1115 Spanish verbs, we obtained 33 new German verbs. This

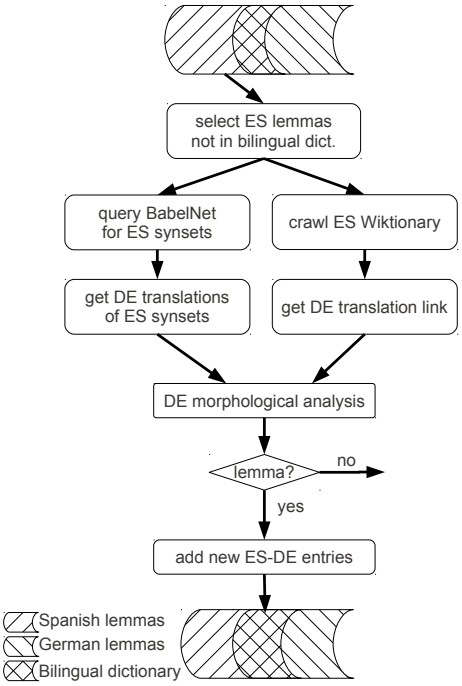

Figure 3: Compilation workflow

results in a recall of under 3%, which shows the limit of the method.

In a next step, we measured the overlap for the nouns[5] before and after harvesting the BabelNet translations: the 594 newly covered nouns represent an increase of 8%. The following examples of missing word equivalences show that we can manually find their German translations: *abigeo* (de: Viehdieb; en: rustler, cattle thief), *zapallo* (de: Kürbis; en: pumpkin). However, we want to translate as many of these words as possible automatically into German. Looking at the failures, we observe a large number of participles and adjectives analyzed as common nouns. In a next step, we need to loosen the part-of-speech restriction we have imposed on the filtering.

3.5 Corpus coverage

We have collected articles from an online newspaper[6] in order to test the coverage on a Peruvian corpus. This small ad hoc corpus contains about

[5]Note that the "noun" entries in the Spanish-Quechua dictionary also cover Spanish adjectives as there is no morphological distinction between nouns and adjectives in Quechua.

[6]`http://diariodelcusco.com`

10,000 words. In the near future, we will gather more articles and periodically measure the coverage of the growing collection. For the evaluation, we let the MT system do a raw translation (lexical transfer) without lexical disambiguation. Before the extension of the dictionary, the "out-of-vocabulary" ratio of common nouns was 11.95% for tokens and 16.66% for types. With the additional entries extracted from BabelNet, OOV ratios decreased to 7.39% and 11.16%, respectively. Note that the unknown types not only contain single lemmas but also multiword expressions that are not yet listed in the bilingual dictionary.

Applying the same procedure as described in section 3.4, we have added 640 new entries for adjectives to our dictionary. As a result, the OOV ratios of adjective types have decreased from 41.62% to 37.03%. Although the corpus coverage for adjectives improved, it is still low, partly due to the fact that we have not yet treated the participles as adjectives. For example, our dictionary does not have adjective entries for common verb participles like *acompañado* (en: accompanied). Other examples of untranslated adjectives are some toponyms like *limeño* (from Lima), missing from our bilingual dictionary, and *cusqueño* (from Cuzco), absent even from the Spanish full form lexicon. Some common adjective pairs might not be found in BabelNet, e.g. *virtual - virtuell*, but are present in the Wiktionary, and vice versa. For this reason, we combined all possible sources in order to maximize the automatic extension of our dictionary.

4 German morphology features

Apart from extending the dictionary with new entries, we added the missing parts of the morphological information needed for the translation from Spanish to German.

4.1 German noun gender

For German nouns, in addition to the lemmas, we need at least the gender. In fact, the minimum information depends on the morphological tool we use to generate the German forms.[7] Due to the German agreement constraints, we need the gender of a noun in order to generate the correct inflections on the elements of the noun phrase.[8] Gender information is unequally present in the different sources we have exploited: Almost all the entries retrieved from the "Ding" lexicon and the Wiktionary pages contain the gender of the noun, but BabelNet does not indicate this information.

We applied the same morphology tool (Clematide, 2008) used for generation to analyze the German side of the –with respect to the gender– underspecified dictionary entries. We extracted the analyses with more than one possible gender and manually checked whether the selected gender corresponded to the intended meaning of the Spanish-German lemma pair. We observe different kind of ambiguities: There are true gender alternatives, e.g. *der/das Hektar* is both masculine and neuter, but also homographs with different senses: *die Flur* (en: acre) vs *der Flur* (en: hall). Variable word segmentation within compounds leads to another type of gender ambiguities: the feminine derivative *die Wahrsagerei* (en: fortune telling) is more probable than the neuter compound *das Wahrsager-Ei* (en: the fortune teller's egg).

Automatic gender attribution through morphological analysis is error-prone and far from complete. Nearly a third of the candidate entries extracted from BabelNet received an analysis. We have manually annotated 5% of those entries to roughly estimate the a posteriori precision: 78.5% are correct, 16% wrong, and about 5.5% unclear.

Finally, we need to include the linguistic gender alternation paradigm to gentry nouns and professions. For example, the Spanish word *periodista* refers to both the male and female journalists, but German distinguishes between *Journalist* (masc.) and *Journalistin* (fem.).

4.2 German verb auxiliary

German verbs typically use only one of the two auxiliary verbs –*haben* or *sein*– to form the perfect tenses. Nevertheless, some verbs may alternatively use one or the other, depending on the context. Reflexive verbs never use the auxiliary *sein* nor do verbs with a direct object. The most common verb type that requires *sein* as auxiliary are motion verbs, such as *fahren* (en: drive). But if the same verb[9] has a direct object, the auxiliary *haben* appears in the perfect tense form.

[7] This would also be necessary for Spanish, but we are translating only in one direction, from Spanish to German.

[8] Note that German adjectives are inflected according to the gender of the head noun, e.g. in accusative case '*die **grosse** Frau*' (the tall woman) vs '*den **grossen** Mann* (the tall man).

[9] The same surface form may have different verb subcategorization frames.

sein: *Ich bin von A nach B gefahren.*

(1) *Ich* ***bin*** *von A nach B gefahren.*
I **am** from A to B driven.
"I drove from A to B."

haben: *Ich habe [mein Auto]$_{DirObj}$ von A nach B gefahren.*

(2) *Ich* ***habe*** *mein Auto von A nach B gefahren.*
I **have** my car from A to B driven.
"I drove my car from A to B."

Where do we get this information from and how should we best encode this alternative behavior in our dictionary? Unfortunately we cannot automatically get the auxiliaries for every German verb from Canoo, so we extracted 4056 verbs from the Wiktionary dump made available by Henrich et al. (2011). Furthermore, we collected 5465 pages by crawling the Wiktionary for German verbs[10]. As Tab. 2 shows, there are more verbs with auxiliary *haben* than with *sein*, therefore we choose the auxiliary *haben* to be the default. We filtered the verbs with *sein* from both sources and merged them, which resulted in a list of 394 verbs[11].

Source	verbs	auxiliaries		
		haben	sein	both
dump2011	4056	3721	293	17
crawl2013	5469	4814	351	200
merged			394	

Table 2: Auxiliary verb distribution

The header of our dictionary contains a specific paradigm for the verb entries for which the German translation has to be generated with *sein* in the perfect tenses. This is a derivative version of the default verb paradigm, as Fig. 4 shows.

To select the correct auxiliary we need the syntactic analysis of the German verb phrase or at least the information about the presence or absence of a direct object. If the parse tree obtained from the analysis of the Spanish source sentence is erroneous, we must rely on other means to disambiguate the verb auxiliaries. Which methods are best suited to solve this task is a topic for future work.

[10] `http://de.wiktionary.org/w/index.php?title=Kategorie:Verb_(Deutsch)` [retrieved 2013-12-27]

[11] 43 verbs are only in dump2011, 101 only in crawl2013, 250 in both lists.

```
<pardef n="VM_VV_MAIN_BE">
 <e>
  <p>
   <l><s n="parol"/>VM</l>
   <r><s n="aux"/>sein<s n="pos"/>VV</r>
  </p>
  <par n="Verb"/>
 </e>
</pardef>
```

Figure 4: Paradigm definition (`<pardef>`) for main verb pairs (es:`VM`–de:`VV`) with explicit value *sein* for the auxiliary attribute (`aux`) on the German side (`<r>`).

5 Conclusion

In our hybrid MT system with a rule-based kernel, the bilingual dictionary plays a crucial role. We have built a Spanish-German dictionary from different freely available resources with general MT in mind. This dictionary contains around 33,700 entries at the moment of writing.[12]

This paper describes the extraction of new entries from BabelNet and Wiktionary. We have shown that these sources can both contribute to the enhancement of our dictionary, albeit on different scales and in a complementary manner. Encouraged by the coverage boost yielded from the addition of nouns and adjectives extracted from BabelNet, we want to apply a similar procedure to verbs. We will also crawl the Wiktionary for the Spanish adjectives and their German equivalents, and continue to gather more information from the net as it gets available. Word derivation is another issue that we want to address, mainly to cover adverbs with the suffix *-mente*, and also to include even more adjectives.

Acknowledgments

The author would like to thank Annette Rios for her helpful advise and for proof-reading the final version of this paper. This work is funded by the Swiss Nation Science Foundation under grant 100015_132219/1.

[12] Available from our project's website: `http://tiny.uzh.ch/2Q`

References

Bogdan Babych, Kurt Eberle, Johanna Geiß, Mireia Ginestí-Rosell, Anthony Hartley, Reinhard Rapp, Serge Sharoff, and Martin Thomas. 2012. Design of a hybrid high quality machine translation system. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 101–112, Avignon, France, April. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors. 2012. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Simon Clematide. 2008. An OLIF-based open inflectional resource and yet another morphological system for German. In A. Storrer, A. Geyken, A. Siebert, and K. M. Würzner, editors, *Text Resources and Lexical Knowledge*, number 8 in Text, Translation, Computational Processing, pages 183–194. Mouton de Gruyter, Berlin, Germany, September. KONVENS 2008: Selected Papers from the 9th Conference on Natural Language Processing.

Andreas Eisele and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), May.

Carla Parra Escartín. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In Calzolari et al. (Calzolari et al., 2012).

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*, Matsue, Japan.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2011. Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. In *Proceedings of the 5th Language & Technology Conference (LTC 2011)*, pages 126–130, Pozna, Poland, November.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit*, pages 79–86, Phuket, Thailand, September. European Association for Machine Translation.

Montserrat Marimon, Natalia Seghezzi, and Núria Bel. 2007. An Open-source Lexicon for Spanish. *Procesamiento del Lenguaje Natural*, 39:131–137.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2012. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, (25):53–82.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In Calzolari et al. (Calzolari et al., 2012).

Annette Rios and Anne Göhring. 2012. A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank. In Calzolari et al. (Calzolari et al., 2012), pages 1874–1879.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Marroco.

Špela Vintar, Darja Fišer, and Aljoša Vrščaj. 2012. Were the clocks striking or surprising? Using WSD to improve MT performance. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 87–92, Avignon, France, April. Association for Computational Linguistics.