Design and compilation of a specialized Spanish-German parallel corpus

Carla Parra Escartín

University of Bergen
Bergen, Norway
carla.parra@uib.no

Abstract

This paper discusses the design and compilation of the TRIS corpus, a specialized parallel corpus of Spanish and German texts. It will be used for phraseological research aimed at improving statistical machine translation. The corpus is based on the European database of Technical Regulations Information System (TRIS), containing 995 original documents written in German and Spanish and their translations into Spanish and German respectively. This parallel corpus is under development and the first version with 97 aligned file pairs was released in the first META-NORD upload of metadata and resources in November 2011. The second version of the corpus, described in the current paper, contains 205 file pairs which have been completely aligned at sentence level, which account for approximately 1,563,000 words and 70,648 aligned sentence pairs.

Keywords: corpus compilation, specialized parallel corpora, Machine Translation

1. Introduction

In this paper, the design and compilation process of a specialized Spanish-German parallel corpus is described. Gale and Church (1991), Dagan et al. (1993), Kupiec (1993), Fung (1995), Smadja et al. (1996), Melamed (1997), Tiedemann (1998), Tufis (2002) and many other researchers have discussed the importance of parallel and comparable corpora. In the particular case of terminology extraction and bilingual lexica induction, this need for corpora is even more obvious, as it is in corpora where we can find the necessary linguistic evidence for the identification of translation equivalents of words (Dagan and Itai, 1991; Gale et al., 1992; Kilgarriff, 1999; Piperidis et al., 2000; Brown, 1997, and others). As it is discussed in Melamed (1996), the automatic extraction of bilingual lexica relies heavily on the statistical analysis of parallel corpora and in fact various statistical approaches have been proposed for building these lexica (Brown et al., 1993; Gale and Church, 1991; Hiemstra, 1997; Smadja et al., 1996). In the previously mentioned papers, researchers describe the usage of sentence-aligned parallel corpora to compute the association score between pairs, which enables them to extract correlations of words. Finally, these sentence-aligned corpora are also used to evaluate word correspondences and produce word-level aligned texts (Brown et al., 1993; Gale and Church, 1991).

Moreover, in the last few years the need for domain specific corpora has increased. Many research projects and their publications list specialized corpora among their expected results and/or use the existing corpora because they deal with specialized domains (e.g. FP7 Projects Mormed[1], Pluto[2], Accurat[3], Molto[4], Panacea[5], TTC[6]).

In what follows, the different stages of the corpus compilation process will be explained, as well as the current status of the resource, the different aspects taken into consideration during the design and planning process and both its intended usage and other possible usages.

2. Parallel corpora and Research in Natural Language Processing (NLP)

This section includes two different subsections. In the first one (2.1.), the specific project for which the corpus is being compiled is explained, whereas in the second one (2.2.) other potential usages for the corpus are mentioned.

2.1. Research project related to the project and corpus needs

The research reported in this paper is part of a larger PhD project which aims at improving Spanish-German multiword alignments whenever nominal compounds are involved in German as a source or target language. In fact, complex nominal compounds in the Germanic languages represent a major problem for both human and machine translators when translating between Romance and Germanic languages, such as the Spanish-German language pair.

Thus, not only do the machine translation systems normally fail to translate German compounds into the appropriate Spanish phraseological expressions, but they also fail to produce the German compound nouns a native German translator would suggest, thereby rendering translations in both directions inaccurate. This phenomenon constitutes a great challenge for the induction of bilingual lexica as this is usually based on subsentential alignment (Brown et al., 1993; Vogel et al., 1996) and it has yet not been proven accurate enough as regards to the alignment of multiwords. The main aim of the research project is to improve multiword alignment to single words (1:n alignments) for Germanic and Romance language pairs, starting with a search for latent linguistic clues that may help to automatically identify phraseological expressions in Romance languages that are

[1] http://www.mormed.eu/
[2] http://www.pluto-patenttranslation.eu/
[3] http://www.accurat-project.eu/
[4] http://www.molto-project.eu/
[5] http://www.panacea-lr.eu/
[6] http://www.ttc-project.eu/

likely to be translated into Germanic languages as nominal compounds. Initially, the experiments will be conducted on the Spanish-German language pair, but the project will be expanded to include the Norwegian-Spanish language pair in later stages.

Within the experiments done with the corpus it will be tested whether the terms referring to the main topic of a text are potential candidates to produce compounds in German. Should this hypothesis be proven true, identifying the key words in Spanish and their translation into German or using domain specific lists would constitute a very valuable clue towards the identification of Spanish phraseological expressions and their compound correspondences in German. Furthermore, the research project also aims at testing whether domain tuning is actually necessary to extract correspondences and thus the 10 domains crawled so far will enable to carry out experiments as regards to this matter. The preliminary assumption is that this will be the case (i.e. domain tuning is necessary), since previous research in Machine Translation has proven precisely that (Koehn, 2002; Koehn, 2010).

2.2. Other potential uses of the corpus

Even though the corpus described here will be used for one specific purpose, efforts have been done as regards to interoperability and standardization of the resource. The main idea behind this effort is that once the corpus is finished other researchers may use it in other projects. As previous efforts to compile parallel corpora have shown (Koehn, 2005; Steinberger et al., 2006), parallel corpora may be used for a wide variety of research purposes within the NLP field. Concretely, parallel corpora may be used for information retrieval, word sense disambiguation, anaphora resolution, induction of tools across languages, training of statistical Machine Translation Systems (as exemplified, for instance, in Koehn (2005)), multilingual categorization, extraction of domain-specific terminology lists, testing and training of document classification software and automatic indexing systems, etc.

3. German-Spanish available corpora

Even if it can be generally assumed that there is currently a need for bilingual specialized corpora in different language pairs, their existence is rather limited and most of the available specialized corpora include English as one of their languages. Thus, working with other language combinations, such as in our case German-Spanish (or later on Norwegian-Spanish), becomes a problem when one needs a specialized corpus. With respect to available German-Spanish corpora aligned at sentence level, the Linguistic Data Consortium[7] offers two written resources with this pair of languages: *The ECI Multilingual Text*[8] and the *Web 1T 5-gram, 10 European Languages Version 1*[9]. However, none of them can be considered a bilingual corpus aligned at sentence level. In the repository of the European Project CLARIN[10] there was only one written resource with these two languages: *The Copenhagen-Dependency-Treebank Project*[11]. However, they are both part of bilingual corpora aligned to Danish and the resource is currently under construction and not available[12]. Finally, the European Language Resources Association Catalogue[13] and its Universal Catalogue[14] were also checked. In ELRA Catalogue there were several written corpora, but some were not having German-Spanish aligned texts, some where just multilingual collections and none of them was really domain specific. The Universal Catalogue lists the most currently used corpus for many language pairs in which one of the languages is English: the *Europarl Corpus*[15]. However, the language pair German-Spanish is not available for download and has to be self-compiled from the individual languages files. The Universal Catalogue also lists the *Acquis Multilingual Parallel Corpus*[16]. This resource would actually constitute the only available large Spanish-German corpus aligned at sentence level, together with the *DGT Multilingual Translation Memory of the Acquis Communautaire*[17]. However, both of them were disregarded. In the case of the *Acquis Corpus*, it was disregarded because the alignment has not been manually corrected, which would imply that the corpus has to be "cleaned" before being used, and because even though it has been domain-tagged with the EUROVOC domains, the original language of each of the files is not specified and that information was also needed. In the case of the *DGT Translation Memory*, the problem was that even though the alignments are of a better quality (they have been corrected), the original texts cannot be reconstructed and therefore it can be considered a big database with a mixture of all possible domains without any possible reordering or document retrieval. Finally, another interesting corpus was found: the *Computer-domain Corpus*[18]. However, this corpus is just restricted to one domain and it had been established to have more than one.

As can be concluded from the previous explanation, there is a clear lack of domain specific corpora aligned at sentence level for the German-Spanish pair of languages. This is actually not surprising, since corpus compilation is a time consuming task and therefore uncommon language pair combinations do not usually have available corpora, whereas very commonly researched language pairs (e.g. English-French or English-Spanish) have a greater number of resources available.

Since a corpus to carry out the experiments in the research project was needed, it was decided to compile a specialized German-Spanish corpus that could be also made available

[7] http://www.ldc.upenn.edu/Catalog/

[8] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC94T5

[9] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T25

[10] http://catalog.clarin.eu/ds/vlo/

[11] http://www.clarin.eu/node/3296

[12] http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT

[13] http://catalog.elra.info/

[14] http://universal.elra.info/

[15] http://www.statmt.org/europarl/

[16] http://langtech.jrc.it/JRC-Acquis.html

[17] http://langtech.jrc.it/DGT-TM.html

[18] http://universal.elra.info/product_info.php?products_id=398

for other research projects. The next section provides further details about the corpus itself.

4. Description of the TRIS corpus

The first step as regards to compiling a corpus is to establish where to obtain the data from. As it is commonly known, international organizations offer us the best possible environment to collect multilingual resources because they usually translate all their official publications into the languages of their member states (i.e. EU bodies) or into the languages that have been stablished as their official ones (i.e. UNO). As it has previously been stated, the aim was to compile a corpus for the pair of languages German-Spanish using specialized texts from a variety of domains. Since it was also desired to have information about the original source language, it was finally decided to compile a corpus with texts coming from the European Commission and more concretely from the DG Enterprise and Industry Project[19]. This choice was made because it ensured that the corpus would not only have specialized texts in different languages, but also information about the domain, subdomain, country of origin, year of publication, etc. In what follows further details are given as regards to the 98/34/EC Directive (4.1.), the TRIS database (4.2.) and the copyright (4.3.).

4.1. The 98/34/EC Directive (TRIS)

As stated in their website,[20], *"the 98/34/EC Directive (formerly 83/189/EEC) sets up a procedure which imposes an obligation upon the Member States to notify to the Commission and to each other all the draft technical regulations concerning products and soon Information Society Services before they are adopted in national law. Such procedure aims at providing transparency and control with regard to those regulations. Since they could create unjustified barriers between Member States, their notification in the draft form and subsequent evaluation of their content in the course of the procedure help to diminish this risk."*
This obligation to submit any national draft technical regulations to the European Commission and the fact that all those regulations are translated into the languages of all member States constitutes a very valuable source of new data for research in Natural Language Processing and linguistics related with specialized domains. This is because technical regulations, notwithstanding their legal nature, include a lot of technical and specialized terminology in the relevant fields and can be also considered technical texts as regards to the contents they deal with and the way they are written.

4.2. The TRIS database

The TRIS database is publicly available and the user is provided with a search interface in which the search can be tuned according to the specific needs. Figure 1 shows the search interface and how the user can select one or several features to find the appropriate files in the database.
The "Product Type" attribute refers to the different thematic domains and subdomains. This is possible because

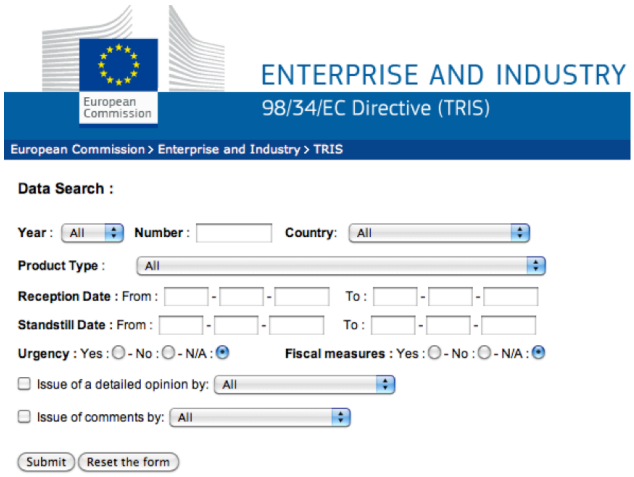

Figure 1: Screen capture of the search interface of the TRIS database

every file is classified into a domain and a subdomain if it is the case. The project covers a very wide range of thematic areas, subdivided in different related topics (see Table 1 below to gather an idea of the main domains available in the database). Furthermore, as texts may be filtered according to their country of origin (attribute "Country" in the search interface, the crawler was defined in such a way that it would only crawl texts produced in Germany or Austria in the case of German and in Spain in the case of Spanish and which had been translated into the other language (the translations can only be retrieved at document level). Finally, only complete years at the date of crawling (June 2011) were crawled and therefore the corpus only contains texts from 1990 (the first available year in the database) to 2010. Future years may however be included on the long run.
As may be concluded, the final corpus is not only divided into different domains (and subdomains) and covers a wide temporal frame (20 years; 1990-2010), but also includes two different subsets: Spanish > German translations and German > Spanish translations (the German texts can be divided into those written in Austria and those written in Germany). The table 1 also shows the number of original source files crawled for each domain and country of origin. Note that even though the database comprises 13 different domains, no files – from Germany, Austria or Spain – were found for the last three in the list (T00T: Transport; V00T: Telecoms; and X00M: Goods and Miscellaneous Products) and therefore are not present in the corpus described here.

4.3. Legal Notice

As it is also pointed out in Koehn (2005), *"usually, there are also copyright concerns, although less so for information from government sources"*. Similarly to what happens with the texts from the European Parliament, the legal notice stated at the website of the European Commission[21] also indicates that *"Reproduction is authorised, provided the source is acknowledged, save where otherwise stated"*.

[19] http://ec.europa.eu/enterprise/tris/index_en.htm
[20] http://ec.europa.eu/enterprise/tris/about/index_en.htm
[21] http://ec.europa.eu/geninfo/legal_notices_en.htm

Field	No. files Austria	No. files Germany	No. files Spain
B00: Construction	205	174	39
C00A: Agriculture, Fishing and Foodstuffs	52	60	78
C00C: Chemicals	16	19	12
C00P: Pharmaceuticals and Cosmetics	3	17	3
H00: Domestic and Leisure Equipment	12	7	36
I00: Mechanics	28	8	45
N00E: Energy, Minerals, Wood	22	14	14
S00E: Environment	24	27	12
S00S: Health, Medical Equipment	4	1	2
SERV: 98/48/EC Services	15	38	9
T00T: Transport	0	0	0
V00T: Telecoms	0	0	0
X00M: Goods and Miscellaneous Products	0	0	0

Table 1: Main domains included in the corpus and number of files crawled per country of origin in each domain

Since the TRIS database does not have any statement which indicates the opposite, to the best of the author's knowledge the corpus compilation is not against any copyright laws as long as the source is explicitly acknowledged.

5. Corpus compilation process

After determining the kind of corpus to be compiled, there are other important issues to be taken into account such as the general roadmap for the corpus compilation, the format of the metadata schema, corpus encoding, etc. The following subsections briefly explain how the corpus compilation was planned, which stages are being done and what is already finished.

5.1. General outline and project stages

Corpus compilation is known to be a slow and time consuming process and therefore it shall be well planned beforehand. Figure 2 shows the different stages that were envisaged for the compilation of the corpus.

As can be seen in the figure, four distinct stages have been established. In the first one, all relevant documents were automatically crawled, and in the subsequent second stage all crawled files were classified by country of origin, domain and year of publication as well as paired with the corresponding translation in the target languages (German/Spanish). During this stage, all files which were not in the target languages of the project were disregarded (sometimes a file tagged as Spanish or German was suddenly in French or Italian, for example). Besides, other files had to be disregarded as well due to the fact that their format was either corrupted or not compatible with current software versions, which made it impossible to open them. These files were mainly the oldest ones. Concretely, as regards to Austria and Germany files prior to 1999 had to be excluded, and in the case of Spain those prior to 1997. As a result of

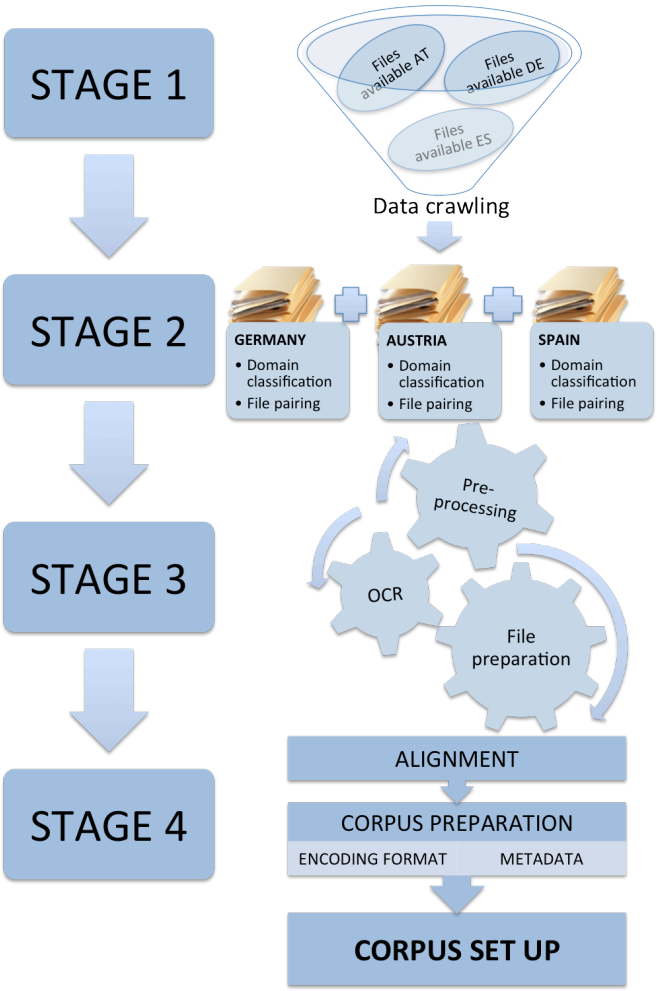

Figure 2: Corpus compilation stages

this second stage, the corpus consists of 995 file pairs (380 from Austria, 365 from Germany and 250 from Spain). In table 1 the distribution of files across domains can be observed. Thus, both Austria and Germany coincide in having most of their files in the domains Construction (B00) and Agriculture, Fishing and Foodstuffs (C00A), whereas the majority of the Spanish files belong to Agriculture, Fishing and Foodstuffs (C00A) and Mechanics (I00).

The third corpus-compilation stage deals with all the preprocessing steps required prior to alignment. All scanned documents as well as all non-editable pdf-documents were identified and converted to an editable format (rtf, to be more precise) by means of an Optical Character Recognition (OCR) system. OCR processed documents undergo then a human supervision and are manually corrected to ensure that the alignment process will not encounter any format problems and that the text is appropriately written (i.e. spelling errors produced by the OCR system are also corrected). Right now, all pdf files corresponding to texts written in Austria have been processed and are now being aligned. Files coming from Germany have been partially finished and Spanish files have not been supervised yet. Although this process is a simple task, it is time consuming as one has to read the original pdf file and the converted file

and correct all mistakes.

As far as non-scanned files are concerned, they also undergo a human supervision as there are sometimes pages, paragraphs or sentences missing in the translation/original file and this would produce a lot of noise in the alignment stage. The original files are edited to ensure that both the source and target language files are "mirrors" of each other. This process is relatively quicker than the one involving OCR files supervision. Nonetheless, sometimes some minor formatting/editing corrections have to be made. Currently all the files from Austria have been preprocessed and are being aligned, while files from Germany and Spain are currently still being pre-processed.

Even though there is usually a tokenisation and sentence splitting step prior to sentence alignment, in the case of the corpus described here this was not done. This was already discussed by Koehn (2005) in relation with the compilation of the Europarl Corpus. In his case tokenization and sentence splitting were disregarded because they are error prone processes and require specialized tools for each language. As regards to the corpus described here, these were not the reasons behind this decision, but rather the fact that a different approach was established since the original files were MS Word files and therefore it would have been necessary to convert all files to plain text before undergoing the tokenization and sentence splitting steps.

The fourth and last stage prior to the final corpus release starts after the files are ready for alignment at sentence level. The commercial software SDL Trados and concretely its alignment tool (WinAlign) is used as it allows for sentence alignment from native MS Word files and no format conversion prior to alignment is required. The decision was taken due to practical reasons: WinAlign saves time at this stage of the process while producing bilingual files than can easily be converted to the standard Translation Memory eXchange (TMX). This is currently done by means of a perl script.

Figure 3 shows the user interface of WinAlign. As can be seen, the program proposes automatic alignments (dotted lines), and a human validator can correct those alignments, confirm (line) or reject them (no line at all). The program also permits the user to join or split segments as well as edit them if needed. This is very useful as sometimes it is necessary to join several segments. This is the case, for example, when in the original MS Word file in German there is a list with the verb in a separate line at the end of the list while in the Spanish translation the verb occurs at the beginning of the list. German grammar requires that certain structures have the verb at the end and this cannot be done in Spanish. The editing feature of WinAlign allows the user to edit those segments and join/split them accordingly so that they are paired with the appropriate Spanish sentence. Besides, since text editing is also possible, if any minor mistake is detected while supervising the automatic alignment proposed by WinAlign it is easily corrected.

Figure 4 shows the structure of an aligned segment produced by SDL Trados WinAlign. As can be acknowledged, its tagged structure makes it feasible to automatically convert all aligned files into the common TMX standard. To this end, a perl script was created that automatically processes all files and converts them to TMX. Figure 5 shows

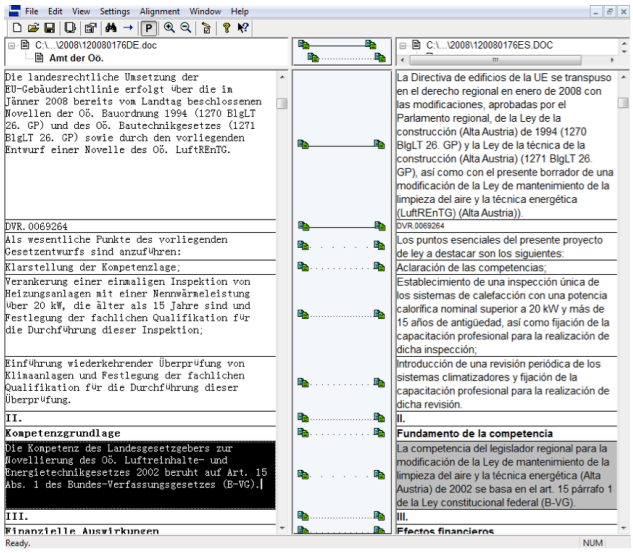

Figure 3: The SDL Trados WinAlign Interface

the structure of the TMX files produced by the perl script.

```
<TrU>
<Quality>100
<CrU>ALIGN!
<CrD>06032012, 19:05
<Seg L=DE-AT>Die ÖNORM EN 1317 stellt das Grundsatzregelwerk zum
Verständnis dieser RVS dar.
<Seg L=ES-ES>La norma ÖNORM EN 1317 es el aparato normativo
básico para la comprensión de estas RVS.
</TrU>
```

Figure 4: Aligned segment produced by SDL Trados WinAlign

Once all files have been processed, the bilingual files will also be split in monolingual files. Thus, the corpus will be released in several formats, and researchers will be able to select the appropriate format for their research purposes.

120030038.tmx

```
1  <?xml version="1.0" encoding="UTF-8"?>
2      <!DOCTYPE tmx PUBLIC "-//LISA OSCAR:1998//DTD for Translation Memory
   eXchange//EN" "tmx14.dtd" >
3      <tmx version="1.4">
4
5      <header
6          creationtool="SDL Trados WinAlign 8.3.0.863"
7          creationtoolversion="Edition 8 Build 863"
8          o-tmf="SDL TM8 Format"
9          segtype="sentence"
10         adminlang="EN-US"
11         srclang="DE-AT"
12         datatype="xml"
13         creationdate="February 2012"
14         creationid="Carla Parra, UiB"
15     >
16     </header>
17
18     <body>
19
20 <tu tuid="B00Y2003S1" creationdate="20120126T2111Z" creationid="ALIGN!">
21 <tuv xml:lang="DE-AT">
22 <seg>STRASSENERHALTUNG</seg>
23 </tuv><tuv xml:lang="ES-ES">
24 <seg>MANTENIMIENTO DE CARRETERAS</seg>
25 </tuv>
26 </tu>
27
28 <tu tuid="B00Y2003S2" creationdate="20120126T2111Z" creationid="ALIGN!">
29 <tuv xml:lang="DE-AT">
```

Figure 5: Sample of a TMX aligned file

5.2. Corpus encoding

The corpus will be released in several encoding formats. On the one hand and as it has been stated before, it will

be released in TMX since this a relevant standard in our field and these files can easily be converted into the format required by Machine Translation Systems like MOSES [22]. Besides, there will be another release encoded in XML TEI P5 (Sperberg-McQueen and Burnard, 2009) as this is another important standard in our field.
All files which undergo Part-of-Speech tagging (PoS-tagging) will also be released. Although the decision is not definite yet, most probably PoS-tagging will be done using the tree-tagger tool[23] for annotating texts with part-of-speech and lemma information.

5.3. Metadata schema

Metadata are a key source of information about any language resource or tool and not only provide valuable information about it, but also enhance their visibility and reusability. For this reason, from the very beginning of the corpus compilation process metadata were taken into consideration.
The first release of the TRIS corpus, which was released in the first META-NORD upload of metadata and resources in November 2011, is already described with a metadata schema. This can be seen in Figure 6.
Subsequent releases of the corpus until its final release will also include a Metadata Description compliant with the one provided by Meta-Share and used by its cooperating project META-NORD. As stated in Deliverable 7.2.1 of the META-NET project (Gavrilidou et al., 2011), the metadata schema developed by META-SHARE is based, on the one hand on the results of a user requirements survey carried out by means of interviews during last LREC 2010 and, on the other hand, on the overview of metadata schemas and catalogs. Since the Meta-Share metadata schema is still not definite, the current description is subject to modifications in the near future.

5.4. Corpus licensing and availability

As soon as the corpus is finished, it will be made publicly available so that this compilation effort can contribute towards the development of research in the NLP field. As explained in section 2.2., the corpus described here may be used for many other projects besides the specific one also previously described here (see section 2.1. for further details).
As regards to licensing, the corpus is to be made available with restrictions, in particular for research only, whereas a special license will be issued for commercial exploitation. As the resource is being compiled with public funds, the author understands that it should benefit research and also the development of commercial applications. However, mechanisms for handling resources with more than one license type depending of the intended usage of the resource –one implying the payment of a fee for the resource exploitation– are currently not available within the META-NORD network. Cooperation on this issue with META-NORD will be carried out in order to ensure that all interested parties can have access to the resource. As of its current status, a first version has been released within the META-NORD network and is not directly downloadable because licensing is to be compliant with the META-SHARE network and its repository does not offer yet practical technical solutions for handling commercial licenses involving binding agreements, payment, etc.

[22] http://www.statmt.org/moses/

[23] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

6. Conclusions and future work

In this paper the compilation and design process of a specialized bilingual corpus aligned at sentence level has been described and discussed. The different compilation stages have been described with the aim of establishing a corpus compilation methodology for future compilation efforts.
This paper describes an on-going project and the corpus will be released shortly. Due to time restrictions (it is being compiled within a PhD Dissertation project), further work on the corpus (i.e. enlarging it, adding new languages, etc.) is not foreseen although the author would be pleased to cooperate with other researchers who wish to contribute to this effort and help to compile a larger corpus following the guidelines established here.
As has also been previously stated, the main aim is that the corpus is as interoperable and reusable as possible, so that research in our field can profit from this effort.
Possible extensions of this project would be the establishment of a way for correcting OCR files and automatic alignments by means of crowdsourcing (i.e. by using Amazon's *Mechanical Turk*[24]), or mapping the domains and subdomains of the TRIS database with the EUROVOC Subject Domain Classification (Eurovoc, 1995), which would make it more interoperable with other resources, such as the JRC-Acquis (Steinberger et al., 2006).

7. Acknowledgements

The research reported in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405 (project CLARA[25]).

8. References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311, June.

Ralf D. Brown. 1997. Automated dictionary extraction for "knowledge-free" example-based translation. In *In Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 111–118.

Ido Dagan and Alon Itai. 1991. Two languages are more informative than one. In *In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137.

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8.

[24] https://www.mturk.com/mturk/welcome

[25] http://clara.uib.no

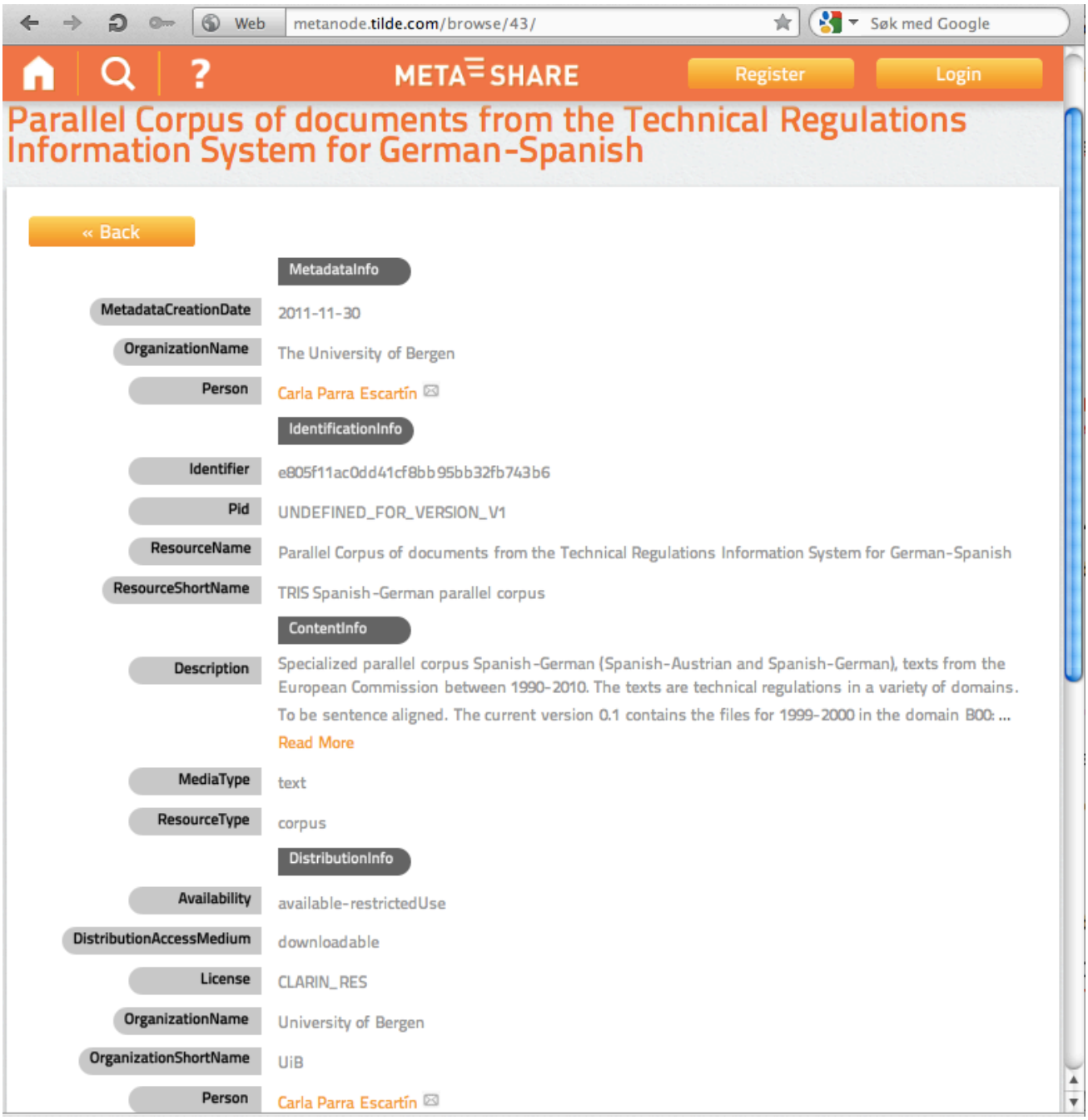

Figure 6: Screenshot of some of the current metadata available for the TRIS corpus at metanode.tilde.com

Eurovoc. 1995. Thesaurus EUROVOC - Volume 2: Subject-Oriented Version. Technical Report Ed. 3/English Language. Annex to the index of the Official Journal of the EC., Office for Official Publications of the European Communities, Luxembourg.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *In Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

William A. Gale and Kenneth Ward Church. 1991. Identifying word correspondences in parallel texts. In *HLT*, pages 152–157.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *In DARPA Speech and Natural Language Workshop*.

Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Manuela Speranza, Monica Monachini, Victoria Arranz, and Gil Francopoulo. 2011. Deliverable D.7.2.1 Specification of Metadata-Based Descriptions for LRs and LTs. Deliverable in the T4ME Project (META-NET.

D. Hiemstra. 1997. Deriving a bilingual lexicon for cross language information retrieval. In *In Proceedings of Gronics'97*, pages 21–26.

Adam Kilgarriff. 1999. "i don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Philipp Koehn. 2002. Europarl: A multilingual corpus for

evaluation of machine translation. Draft.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *In Conference Proceedings: the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 17–22. Association for Computational Linguistics.

I. Dan Melamed. 1996. Automatic construction of clean broad-coverage translation lexicons. In *In Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, pages 125–134.

I. Dan Melamed. 1997. A word-to-word model of translational equivalence. In *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 490–497. Association for Computational Linguistics.

S. Piperidis, H. Papageorgiou, and S. Boutsis. 2000. From sentences to words and clauses. *Parallel Text Processing, Alignment and Use of Translation Corpora*, Text Speech and Language Technology Series:117–138.

Frank A. Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Michael Sperberg-McQueen and Lou Burnard. 2009. TEI P5: Guidelines for electronic text encoding and interchange. Technical report, The TEI Consortium, February.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pages 2142–2147.

Jörg Tiedemann. 1998. Extraction of translation equivalents from parallel corpora. In *In Proceedings of the 11th Nordic Conference of Computational Linguistics NODALI98*.

Dan Tufis. 2002. A cheap and fast way to build useful translation lexicons. In *In Proceedings of the 19th International Conference on Computational Linguistics, COLING2002*, pages 1030–1036.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *In Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.