The English-Slovene ACQUIS corpus

Tomaž Erjavec

Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
Slovenia
tomaz.erjavec@ijs.si

Abstract

The paper presents the SVEZ-IJS corpus, a large parallel annotated English-Slovene corpus containing translated legal texts of the European Union, the ACQUIS Communautaire. The corpus contains approx. 2 x 5 million words and was compiled from the translation memory obtained from the Translation Unit of the Slovene Government Office for European Affairs. The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines TEI P4, where each translation memory unit contains useful metadata and the two aligned segments (sentences). Both the Slovene and English text is linguistically annotated at the word-level, by context disambiguated lemmas and morphosyntactic descriptions, which follow the MULTEXT guidelines. The complete corpus is freely available for research, either via an on-line concordancer, or for downloading from the corpus home page at http://nl.ijs.si/svez/.

1. Introduction

Parallel multilingual corpora are a prime language resource, as they enable induction of translation knowledge in the shape of multilingual lexica or full-fledged machine translation models, and can also serve as welcome aids for translators. The utility of such corpora is even greater if they contain high-quality sentence alignment between the languages, and are linguistically annotated. But parallel corpora, esp. large ones, are hard to produce as a relatively small percentage of text production is translated, and the ever-present problems of copyright and obtaining the digital originals are multiplied and magnified by typically having to involve actors from several countries. However, a large new source of parallel texts is becoming available: many multilingual documents pertaining to the EU are being produced, and, increasingly, made available on the Web. From this source it is then possible to produce and make available highly multilingual corpora. To date, the best known EU corpus is probably Europarl (Koehn, 2002), containing EU parliamentary debates in the 12 languages of the pre-ascension EU states, and having up to 28 million words per language. Still, this corpus does not contain any languages of the new EU or candidate countries, among them Slovene.

In the process of becoming EU member states, candidate countries are obliged to translate core EU legal texts, the so called ACQUIS Communautaire into their own languages. While it is difficult to ascertain which texts exactly constitute the ACQUIS, the approximate size usually quoted is 20,000 pages of text. In Slovenia, the Translation Unit of the Government Office for European Affairs (Službe Vlade RS za evropske zadeve, SVEZ) was responsible for translating the ACQUIS. In order to make the process of translation faster and more consistent, they, similarly to many other translating offices for the ACQUIS, used the Trados Workbench, containing also a translation memory module. Hence, as a side-effect of the translation process, the English-Slovene translation memory for the ACQUIS was built. The government office soon realised the potential use of this resource and accompanying terminological database (Belc and Bratina 2003, Erbič et al. 2005) and made the translation memory, under the name of Evrokorpus, available for searching to the general public at http://www.gov.si/evrokor/, using a custom-built concordance engine (Željko, 2004, 2004a). Furthermore, their multilingual EU terminology database Evroterm was also made available. The concordance engine is also able to heuristically identify the Evroterm terms in the Evrokorpus, thus resulting in useful and much used publicly available service for all translators to and from Slovene.

The corpus presented in this paper was towards the end of 2004 compiled from the source SVEZ English-Slovene translation memory at the Jožef Stefan research institute, IJS. The result, the SVEZ-IJS corpus, is a TEI encoded linguistically annotated parallel corpus of approx. 2 x 5 million words. The terms of the agreement between IJS and SVEZ allow for distribution of the complete corpus for research purposes. The SVEZ-IJS corpus is currently the largest downloadable linguistically annotated parallel Slovene corpus and can thus serve as a useful dataset for a variety of HLT related research tasks.

We should note that a similar corpus to SVEZ-IJS, the JRC-ACQUIS (Erjavec et al., 2005, Steinberger et al., 2006), has recently become available as well. This corpus contains the ACQUIS translated into all the current 20 EU languages and was compiled on the basis of the documents downloaded from the Commission's CELEX and EUR-LEX web pages under http://europa.eu.int/. The greatest difference between the English-Slovene pair of JRC-ACQUIS and the SVEZ-IJS is the fact that the former contains automatically paragraph-aligned full documents, while the latter contains translation memory bi-lingual segments, and is linguistically annotated. This has consequences for the usability of the two corpora: JRC-ACQUIS contains integral and complete texts, but its alignments contain errors, while the IJS-SVEZ contains only separate segments (sentences), but with exact alignments.

The rest of this paper is structured as follows: Section 2 introduces the XML encoding of the corpus; Section 3 details its linguistic annotation; Section 4 discusses the availability and target audience of the corpus; and Section 5 gives conclusions and directions for further research.

```
<ab n="3452">
  <interpGrp resp="svez" type="seg">
    <interp type="status" value="trans" corresp="status.trans"/>
    <interp type="acquis" value="11" corresp="acquis.11"/>
    <interp type="celex" value="32002D0075 32002D0079 32002D0080 32002D0233"/>
  </interpGrp>
  <seg lang="en">
    <w ana="Dd" ctag="DT" lemma="all">All</w>
    <w ana="Ncnp" ctag="NNS" lemma="harbour">harbours</w>
    <c>,</c>
    <w ana="Ncnp" ctag="NNS" lemma="airport">airports</w>
    <w ana="Cc-n" ctag="CC">and</w>
    <w ana="Ncns" ctag="NN">border</w>
    <w ana="Ncnp" ctag="NNS" lemma="station">stations</w>
  </seg>
  <seg lang="sl">
    <w ana="Pg-npn----a" lemma="ves">Vsa</w>
    <w ana="Ncnpn" lemma="pristanišče">pristanišča</w>
    <c>,</c>
    <w ana="Ncnpn" lemma="letališče">letališča</w>
    <w ana="Ccs">in</w>
    <w ana="Afpmpn" lemma="mejen">mejni</w>
    <w ana="Ncmpn" lemma="prehod">prehodi</w>
  </seg>
</ab>
```

Figure 1. An example translation unit from the corpus

2. The encoding of the corpus

The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines TEI P4 (Sperberg-McQueen and Burnard, 2002). In particular, the corpus uses the TEI prose and simple linguistic analysis modules (`TEI.prose` & `TEI.ana`) and is a single XML document (in 35 files) with the `<TEI.2>` root element, which contains the TEI header and the corpus text. The header contains a detailed description of the corpus, including responsibility, availability, editorial policy, information about the source, the tags used in the corpus with the number of their occurrences, etc. Furthermore, the header contains two taxonomies, one detailing the legal status for a translation unit, and the other the legislation area that a translation unit can pertain to. Finally, the header contains the complete list of morphosyntactic codes that are used in the corpus word annotation, represented as TEI feature structure libraries; each such feature structure is also decomposed into its constituent feature-value pairs. We return to this issue in the next Section.

The body of the corpus is composed of a series of translation units (TU); as mentioned, these were taken from the source Trados translation memory, which had been exported in the TMX format. Translation units are encoded as TEI "anonymous blocks", `<ab>` (273,478 occurrences); the structure of a translation unit is illustrated in Figure 1.

The meta-information of the TU is contained within an interpretation group, containing three `<interp>` elements. The first "interpretation" of the translation unit (`type="status"`) gives the legal status of the TU, i.e. which validation steps the translation has undergone; as mentioned, the possible values are detailed in the header taxonomy, and the element is linked with this taxonomy via the `corresp` attribute. The second element (`type="acquis"`) gives the legislation area that the TU is relevant for, e.g. agriculture, institutions, customs, etc. Again, the areas are detailed in the header, and linked to it via `corresp`. The third piece of metadata (`<interp type="celex">`) gives the so called CELEX code(s) that uniquely identify the document(s) that the TU originates from. This information is important for two reasons. First, the TUs in the translation memory, and hence in the corpus, are in document order, so collecting all the contiguous TUs that share the same CELEX code gives an approximation of the complete document, although containing numerous gaps. Second, each CELEX document is indexed by the EU with descriptors from EUROVOC (http://europa.eu.int/celex/eurovoc/), a rich multilingual (meta-)thesaurus covering the fields in which the European Communities are active. So, it would be possible to index each TU with these descriptors, making the corpus useful for document categorisation research (c.f. also Steinberger et al., 2006).

The bi-text of the TU is then encoded in two aligned segments, `<seg>`, one in English and the other in Slovene. The text in both languages is linguistically annotated, and the segments contain word elements `<w>` (10,164,742 occurrences) and punctuation elements `<c>` (1,466,305 occurrences). Words are assigned lemmas and morphosyntactic annotations. Only words where the lemma is different from the actual word-form as it appears in the text contain the `lemma` attribute. The values of `ana` attribute are MULTEXT-East morphosyntactic descriptions (c.f. Section 3.1). The `ana` attribute are pointers (XML IDREF) and link up with their definitions as given in the feature-values in the header. The `ctag` attribute is used only for English, and contains tags from the Penn treebank tagset (c.f. Section 3.2).

3. Linguistic annotation in the corpus

An important feature of the corpus is that it has been pre-processed at the basic linguistic level, namely that of words. This is esp. relevant for the Slovene part of the corpus, as, on the one hand, tools to annotate this language

are not, as for English, widely available, and, on the other, the rich inflectional morphology of Slovene makes using the lemmas instead of the word-forms a very useful option to get around the data scarcity problem ever present in language research.

The linguistic processing of the corpus was composed of

1. tokenisation into words and punctuation
2. part-of-speech tagging, or, rather, word-level syntactic tagging
3. lemmatisation, i.e. computing the base form of the word-forms in the text

For tokenisation we used the mtSeg program (Di Cristo, 1996), which stores the language dependent features in resource files, in particular the abbreviations and split/merge patterns. Both the tagging and lemmatisation were then performed with trainable programs. For these it is, of course, necessary to have the training data-set; in our case this was the Slovene and English part of the MULTEXT-East resources. In the remainder of the section we introduce these learning resources and then detail the word-level syntactic tagging and lemmatisation process.

3.1. The MULTEXT-East resources

The MULTEXT-East language resources, a multilingual dataset for language engineering research and development, first developed in the scope of the EU MULTEXT-East project, have now already reached the 3rd edition (Erjavec, 2004). MULTEXT-East is a freely available (http://nl.ijs.si/ME/V3/) standardised (XML/TEI P4) and linked set of resources, and covers a large number of mainly Central and Eastern European languages. It includes the EAGLES-based morphosyntactic specifications, which define the features describing word-level syntactic annotations; medium scale morphosyntactic lexicons; and annotated parallel, comparable, and speech corpora. The most important component is the linguistically annotated corpus consisting of Orwell's novel "1984" in the English original and translations.

The TEI encoded morphosyntactic features were, for the SVEZ-IJS corpus, taken directly from the MULTEXT-East resources, and became a part of the TEI header for the corpus.

The part-of-speech tagger was trained on the annotated "1984" corpus, and the MULTEXT-East lexicons used to improve its performance. The lemmatiser was likewise trained on MULTEXT-East data for Slovene and English; the complete lexicons for these two languages were used for training.

3.2. The tagging module

For tagging words in the text with their context disambiguated morphosyntactic annotations we used TnT (Brants, 2000), a fast and robust tri-gram tagger. TnT is freely available (but distributed only in compiled code for Linux), has an unknown-word guessing module, and is able to accommodate the large morphosyntactic tagset that is used for Slovene.

The tagger uses two resources, namely a lexicon giving the weighed ambiguity class for each word and a table of tri-grams of tags with weights assigned to the (uni-, bi-, and) tri-grams.

As mentioned, the tagger was both for Slovene and English trained on the MULTEXT-East resources and the results stored in the `ana` attribute of the word tokens. However, unlike Slovene, much larger PoS annotated corpora exist for English than is the small (100,000 words) "1984" corpus. Furthermore, while the MULTEXT-East morphosyntactic annotations for English have the advantage of being expressed in the same formalism as the Slovene ones, they are, unlike the Slovene ones, which have become a de-facto standard for this language, not widely used for English. To offer the users a higher precision tagging and a more familiar tagset, we also tagged the English texts using the TnT model that had been pre-trained on the Penn Treebank, and stored the resulting tagging in the `ctag` attribute (c.f. Figure 1).

While we have not performed a proper evaluation of the tagging error rate, we estimate, on the basis of a small hand-check sample from the corpus, that the per-word accuracy for Slovene is approx. 90%. While this is a relatively low number, it should be noted that most errors are in inflectional features of the word, in particular in the value of inflectional case, which will not have a great impact for the majority of applications.

3.3. The lemmatisation module

Automatic lemmatisation is a core application for many language processing tasks. In inflectionally rich languages, such as Slovene, assigning the correct lemma (base form) to each word in a running text is not trivial, as, for instance, nouns inflect for number and case, with a complex configuration of endings and stem modifications. The problem is especially difficult for unknown words, as word-forms cannot be matched against a morphological lexicon.

For our lemmatiser we used CLOG (Erjavec and Džeroski, 2004), which implements a machine learning approach to the automatic lemmatisation of (unknown) words. CLOG learns on the basis of input examples (pairs word-form/lemma, where each morphosyntactic description is learnt separately) a first-order decision list, essentially a sequence of if-then-else clauses, where the defined operation is string concatenation. The learnt structures are Prolog programs, but in order to minimise interface issues we made a converter from the Prolog program into one in Perl.

For lemmatisation, the text is first tagged, and then, depending on the morphosyntactic description assigned to the word, a Perl function containing the appropriate rule-set is called, which returns the posited lemma of the word. This is then stored in the `lemma` attribute of the word.

The estimated accuracy of the lemmatiser is approx. 95%, i.e. higher than the tagger accuracy. This is due to the fact that, although the lemmatisation depends on the tagging, many errors that the tagger makes do not have any effect on the computed lemma of the word.

4. Availability of the corpus

In order to maximise the usability of the corpus we have concentrated on three facets of availability:

- The corpus is encoded to international standards, in particular XML / TEI P4 and the EAGLES-based MULTEXT(-East) morphosyntactic specifications. An important element of every TEI encoded corpus is

its header, which gives the corpus meta-data. We have made this header available in HTML for easy browsing, with the descriptions of the tags used in the corpus being directly linked into the TEI Guidelines.

- We mounted the corpus under an on-line concordancer, using CQP (Christ, 1994) as the back-end; the Web-based concordancer supports several output views (parallel aligned text, KWIC, word-lists), and enables searches with arbitrary regular expressions over the text and annotations.
- The complete corpus is available for downloading, subject to filling out an on-line agreement, which stipulates that the corpus will be used for research only and that research making use of the corpus will acknowledge the source.

5. Conclusions

The paper has presented SVEZ-IJS, a 10 million word linguistically annotated and standardly encoded parallel corpus, containing ACQUIS Communautaire in English and its translation into Slovene. The corpus is freely available for research, from the corpus home page at http://nl.ijs.si/svez/. We hope that the corpus can serve as a useful dataset for research in Human Language Technologies, esp. those that focus on the Slovene language.

In our further work we plan to concentrate on the following issues:

- As mentioned, the JRC ACQUIS corpus has recently also become available; but while this corpus contains complete texts, its alignments are automatically computed, and thus contain errors. It would be interesting to merge the two corpora, keeping the advantages of both, to arrive at an English-Slovene corpus with integral texts and perfect alignments, which would also be linguistically annotated, and have alignments into further 18 languages.
- In the two years since the production of the SVEZ-IJS corpus the source translation memory of SVEZ has grown considerably, while translations it contains have also been further corrected. We plan to rebuild the corpus from this new translation memory, resulting in an even larger and more authoritative corpus.
- The tagging and lemmatisation models have, in the meantime, also been improved, and could be so even further. In re-compiling the corpus from the new translation memory, the linguistic annotations would thus also be made more accurate.
- Work on other resources for Slovene has also started, e.g. the Slovene Dependency Treebank (Džeroski et al. 2006) and the Slovene WordNet (Erjavec and Fišer, 2006). The SVEZ-IJS corpus could serve as a good empirical basis with which to further improve these new Slovene language resources.

Acknowledgements

The author would like to thank SVEZ for making available their translation memory, in particular Jasna Belc for enabling the transfer. The work on this corpus was supported by the grant CRP V2-0894 and by EU 6FWP projects SEKT and ALVIS.

References

Belc, J. Bratina. S. (2004) From Paper to Collective Knowledge in a "Golden Translation Memory" (Parallel Texts Compilation) and Corpora. *Proceedings of the Slovko conference*. Bratislava.

Erbič, D., Krstič Sedej, A., Belc J., Zaviršek-Žorž, N., Gajšek, N., Željko, M. (2005) Slovenščina na spletu v dokumentih slovenske različice pravnega reda Evropske unije, terminološki zbirki in korpusu (Slovene on the Web in the documents of the Slovene version of the EU law) *Proceedings of the Symposium Obdobja 24: Razvoj slovenskega strokovnega jezika.* Ljubljana.

Brants, T. (2000) TnT-A Statistical Part-of-Speech Tagger. In Proceedings of the 6th Applied Natural Language Processing Conference ANLP-2000 (pp. 224-231). Seattle, WA.

Christ, O. (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*, Budapest, Hungary.

Di Cristo, P. (1996). MtSeg: The Multext multilingual segmenter tools. MULTEXT Deliverable MSG 1, Version 1.3.1. CNRS, Aix-en-Provence.

Džeroski, S., Erjavec, T., Ledinek, N. Pajas, P., Žabokrtsky, Z., Žele, A. (2006) Towards a Slovene Dependency Treebank. *This volume.*

Erjavec, T. (2004) MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*, pp. 1535 - 1538, ELRA, Paris, 2004.

Erjavec, T. and Džeroski, S. (2004) Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence, 18/1 (pp. 17-41). Taylor & Francis.

Erjavec, T., Fišer, D. (2006) Building Slovene WordNet. *This volume*.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005) Massive multilingual corpus compilation: ACQUIS Communautaire and totale. In *Proc. of the Second Language Technology Conference*. April 2004, Poznan.

Koehn, P. (2002) *Europarl: A Multilingual Corpus for Evaluation of Machine Translation.* http://www.isi.edu/~koehn/publications/europarl/

Sperberg-McQueen, C. M. and Burnard, L. (eds.) (2002). *Guidelines for Electronic Text Encoding and Interchange, the XML Version of the TEI Guidelines.* The TEI Consortium, http://www.tei-c.org/

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. (2006) The JRC Collection of the ACQUIS Communautaire – A multilingual parallel corpus with 20+ languages. *This volume*.

Željko, M. (2004) Further use of Terminology and Translation Memory Data. *Conf. on Translating with Computer-Assisted Technology: Changes in Research, Teaching, Evaluation, and Practice.* Rome, Italy.

Željko. M. (2004a) Evroterm and Evrokorpus - Novel Use of Trados Databases. *Proceedings of the Conference Corpus use and learning to translate*. Barcelona.