Erschienen in: Arbeiten zur Mehrsprachigkeit, Serie B 96 (2011) S. 141-144.

A Comparable Wikipedia Corpus: From Wiki Syntax to POS Tagged XML

Noah Bubenhofer, Stefanie Haupt, Horst Schwinn
Institut für Deutsche Sprache IDS
Mannheim
E-mail: bubenhofer@ids-mannheim.de, st.haupt@gmail.com, schwinn@ids-mannheim.de

Abstract

To build a comparable Wikipedia corpus of German, French, Italian, Norwegian, Polish and Hungarian for contrastive grammar research, we used a set of XSLT stylesheets to transform the mediawiki annotations to XML. Furthermore, the data has been annotated with word class information using different taggers. The outcome is a corpus with rich meta data and linguistic annotation that can be used for multilingual research in various linguistic topics.

Keywords: Wikipedia, Comparable Corpus, Multilingual Corpus, POS-Tagging, XSLT

1. Background

The project EuroGr@mm[1] aims at describing German grammar from a multi-lingual perspective. Therefore, an international research team consisting of members from Germany, France, Italy, Norway, Poland and Hungary, collaborates in bringing in their respective language knowledge to a contrastive description of German. The grammatical topics that have been tackled so far are morphology, word classes, tense, word order and phrases. A corpus-based approach is used to compare the grammatical means of the languages in focus. But so far, no comparable corpus of the chosen languages was at the project's disposal. Of course, for all the languages big corpora are available, but they consist of different text types and are in different states of preparation regarding linguistic markup.

Hence we wanted to build our own corpus of comparable data in the different languages. The Wikipedia is a suitable source for building such a corpus. The disadvantage of the Wikipedia is its limitations regarding text types: The articles are (or are at least intended to be) very uniform in their linguistic structure. To overcome this problem we decided to include also the discussions of the articles in our corpus, which can broaden at least slightly the text type diversity.

In this paper we describe, how the Wikipedia was converted to an XML format and part-of-speech-tagged.

2. Wikipedia conversion to XCES

To be able to integrate the linguistic annotated version of the Wikipedia into our existing corpus repository, the data has to be in the XML format XCES[2]. There are already some attempts to convert the Wikipedia to a corpus linguistic usable data source (Fuchs, 2010:136). But they offer either only the data of a specific language version of the Wikipedia in an XML format (Wikipedia XML Corpus, Denoyer & Gallinari, 2006; SW1, Atserias et al., 2008), the format isn't suitable for our needs (WikiPrep, Gabrilovich & Markovitch, 2006; WikIDF, Krizhanovsky, 2008; Java Wikipedia Library, Zesch et al. 2008) or the conversion tool does not work anymore with the current mediawiki engine (WikiXML Collection; Wiki2TEI, Desgraupes & Loiseau 2007). To have a lasting solution, the conversion routines need to be useable also in the future which would allow us to get from time to time a new version of the Wikipedia. Therefore we developed our own solution of XSLT transformations to get an XCES version of the data.

All Wikipedia articles and their discussions are available as mediawiki database dumps in XML (Extensible Markup Language, Bray et al., 1998). These database dumps contain different annotations. Metadata of articles display in XML while the articles display in mediawiki language. We convert these documents into XCES format using XSLT 2.0 transformations to ease research.

[1] See http://www.ids-mannheim.de/gra/eurogr@mm.html.

[2] http://www.xces.org/

This process is divided into 2 sections:

1) The conversion from mediawiki language to XML
2) The conversion from the generated XML to XCES format

The mediawiki language consists of a variety of special signs for special annotations. E.g. to describe a level 2 header the line displays as text wrapped into two equal signs on each side, like this:

```
== head ==
```

Likewise lists display as a chain of hash or asterisk signs, according to the level, e.g. a level 3 list entry:

```
### list entry
```

During the first conversion we process the paragraphs according to their type and detect headers, lists, tables and usual paragraphs. We convert these signs into clean XML, so

```
== head ==
```

turns to

```
<head2>text</head2>
```

and

```
### list entry
```

turns to

```
<item level=3>list entry</item>.
```

Of course inside the paragraphs there may be text-highlighting markup. We access the paragraphs and convert these wikimedia annotations to XML, too. Here we follow a certain pattern to detect text-highlighting signs.

Still the document's hierarchy is flat. In the next step we add structure to the lists. We group the list items according to their level to highlight the structure. In a later step we group all articles into sections depending on the occurrence of head elements. Whenever we add structure we need to take care of possible errors in the mediawiki syntax.

Now the articles need to be transformed into the XCES structure. Here we sort the articles into alphanumerical sections. We transform the corpus and enrich every article with meta data. We provide a unique id for every article and discussion so that they can easily be referenced. Also the actual article text can be distinguished from the discussion part of the article, which is important because they are different text types.

These conversion routines should work for all the language versions of the Wikipedia, but have so far only been tested with the languages necessary for the project: German, French, Italian, Norwegian (Bokmål), Polish and Hungarian.

3. POS-Tagging

To enable searching for word class information in the corpus, the data needs being part-of-speech tagged. This task has not been finished yet, but preliminary tests have been done already. Not having any additional resources, we have to rely on ready to use taggers and cannot do any improvements or adjustments of the taggers.[3] We are using the following taggers:

German TreeTagger (Schmid, 1994) with the available training library for German (STTS-Tagset, Schiller et al., 1995)

French TreeTagger with the available training library for French

Italian TreeTagger with the available training library for Italian

Polish TaKIPI (Piasecki, 2007), based on Morfeusz SIaT (Saloni et al., 2010)

Hungarian System developed by the Hungarian National Corpus team (Váradi, 2002), based on TnT (Brants, 2000)

Norwegian (Bokmål) Oslo-Bergen Tagger (Hagen et al., 2000)[4]

The input for the taggers are raw text files without any XML mark-up and containing only those parts of the Wikipedia, which need to be tagged. So all meta information is being ignored.

A Perl script is used to send the input data in manageable chunks to the tagger. The script also transfers the output of the tagger to a XML file that contains to each token the character position reference to the original data file. Because of the size of the Wikipedia, the tagging process is very time consuming. E.g. the XCES file of the German Wikipedia holds about 15.4 GB of data (785'791'766 tokens). The size of the stand-off file containing the linguistic mark-up produced by the

[3] Nevertheless we get support of the developers of the taggers, which we greatly appreciate.

[4] See `http://tekstlab.uio.no/obt-ny/english/history.html` for the newest developments of the tagger.

TreeTagger (POS information to each token) is about 157.9 GB. It took about 30 hours on a standard double core PC to process this file.

4. Corpus Query System

Our existing corpus management software COSMAS II[5] is used as corpus query system. COSMAS II is currently used to manage the DeReKo (German Reference Corpus, see Kupietz et al., 2010), which contains about 4 billion tokens. Therefore COSMAS II is also able to cope with the Wikipedia data.

To be able to build from time to time new versions of our corpus based on the latest Wikipedia, we can rely on the same version controlling mechanisms as the DeReKo does.

For technical reasons, COSMAS II cannot handle UTF-8 encoding. Therefore the encoding of the XCES files have to be changed to ISO-8859-1 and characters outside this range converted to numeric character references referring to the Unicode code point.

At the end of this process, the Wikipedias in the XCES and the tagged format will be made publicly available to the scientific community.

5. Conclusion

While the Wikipedia is a often used and attractive source for various NLP and corpus linguistic tasks, it is not easy to get an endurable XML conversion routine which produces proper XML versions of the data. It was our attempt to find such a solution using XSLT stylesheets.

After the part-of-speech tagging of the six language versions of the Wikipedia (German, French, Italian, Polish, Hungarian, Norwegian) we are able to build a multilingual comparable corpus for contrastive grammar research in our project.

For future investigations, the advantage of a XML version of the Wikipedia is clearly visible: The XML structure holds all the meta information available in the mediawiki code and can therefore be used to differentiate findings of grammatical structures: Are there variants of specific constructions in different text types (lexicon entry vs. user discussion)? Or does the usage of the constructions depend on topic domains? And how do these observations change in the light of inter-lingual comparisons?

[5] See http://www.ids-mannheim.de/ccsmas2/.

6. References

Atserias, J., Zaragoza, H., Ciaramita, M., Attardi, G. (2008): Semantically Annotated Snapshot of the English Wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation (LREC 08), Marrakech, pp. 2313–2316.

Brants, T. (2000): TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP), Seattle, WA.

Bray, T., Paoli, J., Sperberg-McQueen, C. M. (1998): Extensible Markup Language (XML) 1.0. W3C Recommendation <http://www.w3.org/TR/1998/REC-xml-19980210>.

Denoyer, L., Gallinari, P. (2006): The Wikipedia XML Corpus. In SIGIR Forum.

Desgraupes, B., Loiseau, S. (2007): Wiki to TEI 1.0 project <http://wiki2tei.sourceforge.net/>.

Fuchs, M. (2010): Aufbau eines linguistischen Korpus aus den Daten der englischen Wikipedia. In Semantic Approaches in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2010 (KONVENS 10), Saarbrücken: Universitätsverlag des Saarlandes, pp. 135–139.

Gabrilovich, E., Markovitch, S. (2006): Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In Proceedings of The 21st National Conference on Artificial Intelligence (AAAI), Boston, pp. 1301–1306.

Hagen, K., Johannessen, J. B., Nøklestad, A. (2000): A Constraint-based Tagger for Norwegian. In 17th Scandinavian Conference of Linguistics, Lund, Odense: University of Southern Denmark, 19, pp. 31–48 (Odense Working Papers in Language and Communication).

Krizhanovsky, A. A. (2008): Index wiki database: design and experiments. In CoRR abs/0808.1753.

Kupietz, M., Belica, C., Keibel, H., Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Proceedings of the 7th conference on International Language Resources

and Evaluation, Valletta, Malta: European Language Resources Association (ELRA), pp. 1848-1854.

Piasecki, M. (2007): Polish Tagger TaKIPI: Rule Based Construction and Optimisation. In Task Quarterly 11(1–2), pp. 151–167.

Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R. (2010): Analizator morfologiczny Morfeusz <http://sgjp.pl/morfeusz/>.

Schiller, A., Teufel, S., Thielen, C. (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft, Stuttgart <http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html>.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.

Váradi, T. (2002): The Hungarian National Corpus. In Proceedings of the 3rd LREC Conference, Las Palmas, Spanyolország, pp. 385–389 <http://corpus.nytud.hu/mnsz>.

Zesch, T., Müller, C., Gurevych, I. (2008): Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In Proceedings of the Sixth International Language Resources and Evaluation (LREC 08), Marrakech, pp. 1646–1652 <http://www.lrec-conf.org/proceedings/lrec2008/>.