Analysis of Existing German Corpora

Inna Olifenko and Natalia Borysova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

innaolifenko@gmail.com, borysova.n.v@gmail.com

Nowadays, almost all modern languages have linguistic corpora. The most popular German linguistic corpora, available on the Internet and can be used during the various linguistic studies, are Das Deutsche Referenzkorpus, German political speeches Corpus and Visualization, the NEGRA corpus, the TIGER Treebank, corpus of the Berlin-Brandenburg Academy of Sciences, Limas corpus.

Das Deutsche Referenz korpus (*DeReKo* or *COSMAS II*) [1] is the corpus of contemporary written German of the Institute for German Language in Mannheim. It has 5,4 billion words and constitutes the biggest collection of machine-readable written German. It consists of fiction, science and popular-science texts, a large number of newspaper texts and other written texts. They are constantly extended. DeReKo has 3 subcorpuses: Mannheimer Korpus 1 (mk1), Mannheimer Korpus 2 (mk2) [2], Bonner Zeitungskorpus (bzk) [3]. Mk1 has 293 texts from the period 1950-1967, and about 2,2 million running text tokens. Mk2 has 52 texts from 1949, 1952, 1960-1974, and about 0,3 million running text tokens. Bzk has 10840 texts from 1949, 1954, 1959, 1964, 1969 and 1974, and about 3,1 million running-text tokens.

German political speeches Corpus and Visualization [4] contains the Presidency subcorpus and the Chancellery subcorpus. The Presidency subcorpus has a total of 1442 texts (2392074 tokens), from the period 01.07.1984-17.02.2012. It contains speeches of the presidents: Richard von Weizsäcker (1984-1994), Roman Herzog (1994-1999), Johannes Rau (1999-2004), Horst Köhler (2004-2010) and Christian Wulff (2010-2012). The speeches were got from the online archive of the German Presidency (bundespraesident.de). The Cancellary subcorpus has a total of 1831 texts (3891588 tokens), from the period 11.12.1998-06.12.2011. It contains not only speeches by the chancellors Gerhard Schröder and Angela Merkel), but also a number of other state ministers and a few unrelated speeches of other politicians. The speeches were got from the online archive of the German Chancellary (bundesregierung.de).

The NEGRA corpus [5] version 2 consists of 355096 tokens (20602 sentences) of German newspaper texts. The texts are taken from the Frankfurter Rundschau. The corpus is part-of-speech tagged and completely annotated with syntactic structures. The corpus is stored in an SQL database. Alternatively, the annotations can be represented inline-oriented export format or in PennTreebank format. The different types of information are coded in the corpus: part-of-speech tags (Stuttgart-Tübingen-Tagset (STTS)); morphological analysis (only for the first 60000 tokens, the expanded STTS); the grammatical function in the directly dominating phrase; the category of nonterminal nodes (phrases).

The TIGER Treebank (Version 2.1) [6] consists of app. 900000 tokens

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

(50000 sentences) of German newspaper texts. The texts are taken from the Frankfurter Rundschau. The corpus is part-of-speech tagged and completely annotated with syntactic structures. The annotations can be represented in NEGRA export format or in TIGERXML format. The different types of information are coded in the corpus: part-of-speech tags and morphological analysis (based on STTS, but modified);the grammatical function in the directly dominating phrase; the category of nonterminal nodes (phrases).The TiGer Dependency Bank (TiGer DB) covers 8000-10000 sentences of the TIGER Corpus and is created as a dependency-based gold standard for German parsers. Its annotation is close to the annotation of PARC 700 dependency bank.

Das Wortauskunfts system zur deutschen Sprache in Geschichte und Gegenwart (DWDS-Corpus) [7] is a dictionary digital system, based on very large electronic texts and corpora. It is based on six-volume German dictionary (WDG) and links it with its own texts and dictionaries. This corpus provides the user with information about the correct spelling, pronunciation of sound files and meaning of words with the help of available tags.

Limas corpus [8] contains 500 sources, the total of which has more than 1 million word forms. The collection contains the full texts and passages of different genres. You can not only read sources but also perform keyword search. Today there are three search strategies: simple search, contextual search and phrase search. Data use conditions can be such as text collections can be used for scientific and commercial purposes, provided that citation is done.

In addition to the above corpora, German is processed by such corpora as: The European Parliament hearing corpus, EU documents corpus, InterCorp, Multilingual corpus of the Oslo University, Korpora-Links auf dem Linguistik-Portal LINSE, Lehren un Lernenmit Korpora im DaF-Unterricht and Bibliotheca Augustana, IULA's UPF Textual, plurilingual, specialized Corpus and others.

All considered German corpora certainly have their advantages, but they also have some disadvantages that need be removed.

References

1. Das Deutsche Referenz korpus: [Electronic source]. – Access mode: http://www.ids-mannheim.de/cosmas2/
2. Mannheimer Korpus 1 und Mannheimer Korpus 2: [Electronic source]. – Access mode: http://www.ids-mannheim.de/kl/projekte/korpora/archiv/mk.html
3. Bonner Zeitungskorpus: [Electronic source]. – Access mode: http://www.ids-mannheim.de/kl/projekte/korpora/archiv/bzk.html
4. German political speeches Corpus and Visualization: [Electronic source]. – Access mode: http://perso.ens-lyon.fr/adrien.barbaresi/corpora/index.html
5. The NEGRA corpus: [Electronic source]. – Access mode: http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
6. The TIGER corpus, Treebank and dependency bank: [Electronic source]. – Access mode: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
7. Корпус Берлинской Бранденбургской академии наук: [Electronic source]. – Access mode: http://www.dwds.de/pages/pages_textba/dwds_textba.htm
8. Корпусы института Немецкого языка LIMAS-Korpus: [Electronic source]. – Access mode: http://www.korpora.org/Limas/

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua