Development of Comparable Specialized Corpora of National German Varieties: the UNI-Corpus

Tanja Wissik

University Vienna

Abstract: This paper is dealing with the compilation of comparable corpora to investigate the national varieties of German used in Austria, Germany and Switzerland in the specialized communication in the area of higher education. The comparable corpus is compiled with a special regard to legal and administrative language used in the university system. This paper will present the experience of developing the UNI-Corpus with his three sub-corpora and will discuss the corpus design and issues which arose when setting up the corpus.

1 Introduction

German as a pluricentric language is an interesting research topic, in particular, the national variants in specialized communication. Since previous research dealt primarily with national variants in language for general propose (cf. Ammon 1995) and national variants in LSP were not in the focus of terminological research, the compilation of the UNI-Corpus gives the possibility to study national variants in specialized communication or to describe and contrast national varieties of German in a certain language for special purpose with corpus linguistic methods.

In this paper, the UNI-Corpus is presented and described and some reflections about design criteria for specialized corpora of national varieties are discussed. First, the basic terms relevant for this study are clarified and then some general reflections about specialized corpora of national varieties are discussed before describing the actual design criteria and the

development of the corpus. In the end, some future prospective for exploring this corpus is given.

2 Basic Terms

2.1 Pluricentricity of German

German is not a uniform language as it might seem from some languages classes for foreign language students. Also Markhardt (1993: 3) states that it is not a uniform language and it consists of different varieties. German is a pluricentric language like for example English, Spanish or Portuguese. By pluricentric language is meant, that German is used as an official language in different (parts of) countries like Germany, Austria, Switzerland, Luxemburg, South Tyrol, and Eastern Belgium. That leads to the development of differences in the standard language (cf. Clyne 1992, Clyne 1995). These single linguistically identifiable differences are called language variants. A language variant is "a single unit, or form, as it can be isolated by linguistic analysis from speech or writing" (Ammon 2004: 274). For "a form to be a variant, it has to be an element of a variable, i.e. to be exchangeable paradigmatically (in the Saussure's sense) for at least one other variant" (Ammon 2004: 274). A variety comprises variants and constants and is therefore a system of variants and constants.

2.2 Comparable Specialized Corpora

Before talking about comparable specialized corpora, it has to be clarified what we understand as a corpus. For this paper the definition given by Sinclair (1991: 171) is used where a corpus is defined as a "collection of naturally-occurring language text, chosen to characterize a state or variety of a language" since we are dealing with varieties of language, on one hand with national varieties and on the other hand with language of special purpose (LSP) which is also seen as a variety of language (cf. Adamzik 1997).

As specialized corpora we understand a corpus "that focuses on a particular aspect of a language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers)" (Bowker/Pearson 2002:12) but we do not follow the terminology of these authors, special purpose corpora or LSP corpora when they are restricted to a special domain but apply the terminology used by McEnery et al. (2006). In this study the specialized corpus focuses on a particular domain as well as on particular text types and particular language varieties, so it is in more aspects "special".

McEnery et al. (2006) define a comparable corpus as a "corpus containing components that are collected using the same sampling techniques and similar balance and representativeness […], e.g. the same proportions of the text of the same genres in the same domains in a range of different languages in the same sampling periode" (McEnery et al. 2006: 48). Some other authors only label corpora that consist of more then one language, not containing translations, as comparable corpora (cf. Aijmer 2008, Lemnitzer/Zinsmeister 2010). Also McEnery et al. (2006) exclude corpora containing components of varieties of the same language from the definition of comparable corpora and refer to them as comparative corpora:

> By our definition, corpora containing components of varieties of the same language (e.g. the International Corpus of English […]) are not comparable corpora because all corpora, as a resource of linguistic research, have 'always been pre-eminently suited tor comparative studies' (Aarst 1998), either intralingual or interlingual. The Brown, LOB, Frown and FLOB corpora are typically designed for comparing language varieties synchronically and diachronically. The British National Corpus (BNC), while designed for representing modern British English, is also a useful basis for various intralingual studies (e.g. spoken vs. written, monologue vs. dialogue, and variations caused by sociolinguistic variables). Nevertheless, these corpora are generally not referred as comparable corpora. (McEnery et al. 2006: 48).

In this paper, national varieties are included in the definition of comparable corpora.

As comparable corpus is therefore understood a corpus that consists of more than one languages or more than one language variety, not containing translations, selected and build using the same design criteria and the same sampling techniques.

To sum up, in this study a comparable specialized corpus is a corpus in more than one language or language variety, not containing translations, that focuses on a particular aspect of a language like a particular domain or text type and is build up using the same design and sampling criteria for all the languages or language varieties contained in the corpus.

3 The UNI-Corpus

Below some methodological reflections concerning corpora and the language they contain are discussed (see also Biber et al. 1998) and in particular what does that mean for a specialized corpus of national varieties. Furthermore, the design criteria and the development of the UNI-Corpus are described. The UNI-Corpus is a specialized corpus in the area of university legislation and administration for the national varieties of German. The acronym UNI comes from the topic of the corpus the **uni**versity legislation and administration.

3.1 Representativeness, Balance and Sampling

The UNI-Corpus should represent the German legal and administrative language in the area of higher education. It consists of 3 sub-corpora, one with Austrian texts, one with German texts and one with Swiss texts. Thus, this conception is based on the assumption that we are dealing with three different varieties of German (see 2).

When compiling corpora, one always has to address the issue of representativeness. Representativeness means that “the findings based on its contents can be generalized to the said language variety” (Leech 1991: 27). While talking about representativeness, the type of corpora under question is important. Representativeness for a referential corpus is different to

that for a comparable specialized corpus. But even in specialized corpora the representativeness should not be neglected:

> "Even a specialized corpus, dealing with telephone calls to an operator service should be balanced by including within it a wide range of types of operator conversations (e.g. line fault, request for an engineer call-out, number check, etc.) between a range of operators and customers […] so that it can be claimed to represent this variety of language" (McEnery et al. 2006: 15).

In this work, the corpus is built up according to text types to achieve a representative and balanced corpus (cf. Lemnitzer/Zinsmeister 2010) for the varieties under investigation. More details about the text types are discussed under 3.2.3.

Closely connected to the terms representativeness and balance is the term sampling, especially in relation to referential corpora. "The representativeness of a general corpus depends heavily on sampling from a broad range of genres" (McEnery et al. 2006: 15). But also while building a specialized corpus sampling might get important if the whole population of text of one text type is too big to be included into the corpus. For this corpus, it was the case for the text type study regulation and examination regulation. All the study regulations of all the bachelor, master and PhD studies of all the universities in questions would have been too time consuming and too many to include them all into the corpus. We also assumed after analyzing some bachelor, master and PhD regulations, that there were not many differences in terms of terminology between regulations of different course types within the same university. One option would have been selecting only one type of study program like for example all law program regulations from all universities under investigation. This approach would have excluded the medical universities and the universities for arts. However, in this study no university should be excluded but instead all university specific terminology in the corpus should be included. This permitted on one hand to have a regional distribution by including all universities over the countries under investigation and to individuate national varieties and

on the other hand to filter out specific variants occurring only at one university. The whole population was given by all bachelor programs in all the universities in the all the countries under investigation, only for medicine we allowed diploma courses since medicine has not been converted into the Bologna course model (bachelor and master) and we did not want to exclude the medical universities. After defining the whole population, we chose randomly one bachelor course of each university and its corresponding study and examination regulation.

3.2 Design Criteria and Compilation

3.2.1 Size

An issue that has to be addressed when talking about corpora and corpus design is the size. With the technological advance the corpora are growing faster and faster. There are no rules or norms for the ideal size of a corpus. However, generally specialized corpora have smaller size than reference corpora representing language for general purpose. Indeed, they focus on a special aspect of the language like on special text types or on a specific language for special purpose. So they can not be as big as corpora reflecting the whole language. This statement relies on experience with the texts, on literature (see Pearson 1998: 56; Bowker/Pearson 2002: 45) as well as on theoretical considerations. Since the function of reference corpora is “to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials“ (Sinclair 1996). Therefore it is obvious that it has a much bigger population than a corpus that focuses only one variety or even only one aspect in a variety. Consequently big size should not be seen as the only quality criteria. Regarding corpora representing German national varieties, we have for example the Swiss Text Corpus (CHTK) that has 20 millions of tokens or the whole C4 project where each national subcorpus counts 20 millions tokens. When we look at specialized corpora or LSP corpora, even a 10.000 words corpus can reflect

the variety under investigation (cf. Bowker/Pearson 2002). The UNI-Corpus should represent the German legal and administrative language in the area of higher education in normative texts in the national varieties of German spoken in Austria, Germany and Switzerland. So the corpus will represent two varieties at the same time, a national variety and a certain LSP.

With some search example, it will be illustrated that also with a relatively small but well structured corpus like the UNI-Corpus it is possible to answer relevant questions in the area of university terminology in national varieties of German. Therefore the search results of four randomly chosen terms of university terminology in the DWDS-Corpus, the C4 Corpus and the UNI-Corpus will be compared with each other. In the table below the research results are displayed. The numbers in brackets for example (31) for *Bachelor* shows the absolute frequency of the word, the number below the normalized frequency per 10000 tokens.

Table 1. Comparison of search terms in 3 German corpora

	DWDS	**C4**	**UNI-Corpus**			
			CH	**AT**	**DE**	**total**
Bachelor	(31)* 0,003**	(4) 0,001	(2111) 35,52	(201) 1,54	(2662) 4,49	(4974) 6,38
Bachelor-Studium	(0) 0	(0) 0	(1368) 23,024	(2) 0,02	(73) 0,123	(1443) 1,85
Bachelor-Studiengang	(0) 0	(0) 0	(23) 0,39	(4) 0,03	(340) 0,57	(367) 0,001
Titularprofessor	(2) 0,0002	(10) 0,002	(29) 0,49	(0) 0	(0) 0	(20) 0,03703971

Legends: *absolute frequency ** normalized per 10000 tokens

Now we have a closer look at the concordance lines of the results. From the results in the C4 corpus it seams, that bachelor is not used in Austria, because the results are only from

Switzerland and Germany and not all hits refer to bachelor as a university degree or a university study course.

1. CH 1996 ... in den USA mit einem **Bachelor** of Fine Arts abgeschlossen
2. DE 1986 ... amerikanische Kommilitonen bereits als **Bachelor** of Arts vom College
3. CH 1985 ... Komposition. 1981 « Degree of **Bachelor** of Arts », 1982 « fellowship
4. DE 1928 ... um die Plätze kämpfen. **Bachelors** Quarter - gefällt in der Arbeit,
5. CH 1953 ... Lewinstein, Collinson, Gandolfi, **Bachelor**. Old Boys: Frey; Devick,
6. CH 1953 ... Sitzend: Morris, Levinstein, Collinson, Gandolfi, **Bachelor**. ...

Concordance 1: Bachelor - results in the C4 corpus.

In the DWDS corpus 31 results are found and most of the results refer to the American or English study system. In the Concordance 2 there is a selection of the Results in the DWDS corpus. Furthermore the results are not found in juridical texts and all texts are from Germany. Consequently, with these results, no conclusions can be drawn about the use in Switzerland or Austria.

1 Ge 1906...- engl. Abkürzung für **Bachelor** of arts, d.i. Bakkalaureus…
2 Ge 1906...B.C.L. = **Bachelor** of Civil Law, der unterste Grad der jurist. ...
3 Ge 1906... - B.D. = **Bachelor** of Divinity, in England s.v.w. Kandidat der ...
4 Ge 1906...Baccalaureus medicinae (lat. oder **Bachelor** of Medicine (engl. ,...
5 Ge 1906...Baccalaureus scientiae (lat. oder **Bachelor** of Science (engl. , ...
6 Ge 1906...Medicinae Baccalaureus, engl. **Bachelor** of Medicine, der unterste...
7 Ze 1928...im Gange war, werden um die Plätze kämpfen. **Bachelors** Quarter…
8 Wi 1954...Davies den Grad eines **Bachelor** of Music an der ...
9 Wi 1961... die er 1915 als **Bachelor...**
10 Wi 1973...1925 den Grad eines **Bachelor** of Science und 1927 ...
11 Wi 1973...York den Grad eines **Bachelors** der Sozialwissenschaften.
12 Wi 1973...der Univ. London (**Bachelor** of science 1929, Mus. ...
13 Wi 1973...C. (1906 LL. B. **Bachelor** of Laws). 1907-1912 war ...

14 Ze 1979...betreibt: 1947 erhält sie in Chemie den **Bachelor** of Sciences...

15 Wi 1979...Berkeley den B. Litt. **Bachelor** of Letters), 1907 den ...

16 Wi 1979...einem vierjähr. Kurs zum **Bachelor**' s degree für ...

17 Wi 1989...engl. Komponist, wurde 1606 in Cambridge **Bachelor** of Music...

18 Wi 1995...den USA alle Studien nach dem Bakkalaureus (**Bachelor**), in Deutschland

19 Ze 1998...Der **Bachelor** ist auf drei Jahre ausgerichtet.

Concordance 2: Bachelor - Results in the DWDS corpus

For the search terms *Bachelorstudium* or *Bachelorstudiengang* no results were found in the DWDS corpus.

Below in Concordance 3 a selection of results in the CH-subcorpus of the UNI-Corpus is shown. The search term *Bachelor* was found 4972 times and *Bachelorstudium* 1443 times and *Bachelorstudiengang* 367 times compared to 0 occurences in the other corpora. The search term *Titularprofessor* was only found in the CH-subcorpus, it does not appear in any other subcorpus. So you can draw the conclusion that *Titularprofessor* is a term used only in Switzerland in the area of university terminology.

1 belegt werden, die im **Bachelor** noch nicht besucht worden ist.

2 belegt werden, die im **Bachelor** noch nicht besucht worden ist. d) zu Abfolgen

3 wird mit Einführung der **Bachelor-** und Master- Modul- Modul nach- elemente

4 Auf **Bachelor**-Ebene werden diese Kenntnisse durch Vorlesungen

5 einem späteren Master bildet der **Bachelor** Persisch die Grundlage für eine

6 späteren Master bildet der **Bachelor** Arabisch die Grundlage entfallen gemäss

7 Seminar Orientalisches Seminar **Bachelor** of Arts Arabisch (60 Kreditpunkte)

8 Ziel des **Bachelor** Kredit- punkte benotet weis/e Modultyp ist zum einen

9 belegt werden, die im **Bachelor** noch nicht besucht worden ist.

10 Gesellschaft und Kultur wird mit Einführung der **Bachelor-** und Master-

Concordance 3: Bachelor- some selected Results in the CH-subcorpus of the UNI-Corpus

Based on these reflections and the search example big size should not be seen as the only quality criteria. The UNI-Corpus with approximately 7.8 millions of tokens is a small corpus but it is suitable to answer research questions regarding LSP in the field of higher education and national varieties in combination.

3.2.2 Topic

For general reference corpora topic or subject is not a design criterion, since they should reflect all the aspects of a language. Therefore a wide range of topics should be included. For specialized corpora it is different, even though there are different views. Pearson (1998) states that topic as a design criterion is not very relevant, for Fang (1991) topic is an issue while designing and building a corpus and also Bowker/Person (2002) talk about topic as a design criterion and that sometimes it is hard to delimit the subject. In the end, if topic should be a design criterion for the building of a specialized corpus of national varieties or not, always depends on the purpose of the corpus and the research questions that should be answered and studied with the help of this corpus.

For the UNI-Corpus, the topic is university legislation and administration.

3.2.3 Text Types

As discussed above, the corpus is structured according to text types. Text types were not only chosen to balance the corpus, but also in order to subdivide the field of legal and administrative language in smaller units of analysis. Text types can be defined as classes of texts sharing common typical features according to both linguistic and extra linguistic criteria.

The text can be chosen according to text type classification, if existing, or according to a newly elaborated linguistic and extra linguistic criteria matrix. Regarding juridical and administrative texts, many different text type classifications and text typologies exist (cf. Kjær 1992; Engberg 1993; Busse 1997; Wiesmann 2004). In this paper, the texts are classified

according to Busse's classification (1997: 669ff) regarding the part of the normative texts. As normative we understand according to Busse (1997: 669) actions of persons or institutions in the context of institutional procedures that are going beyond the single case to define in a generalized way what is allowed and what is not allowed. So court sentences are not included in the text type normative texts.

The corpus contains subcorpora of university law, university statutes and study and examination regulations (curricula). Consequently normative texts on national or regional level and also on university level are included in the corpus.

3.2.4 Number of Texts

This design criterion concerns the number of texts of each text type that will be included into the corpus and if these text are from different authors or only from one single author. „In the case of the multi-author corpus, you will be able to get a good idea of what terms and concepts are commonly used in the LSP in question, whereas in the case of the single-author corpus, you will only be exposed to the terms that are preferred by that particular writer" (Bowker/Pearson 2002: 49). As observed by Bowker/Pearson (2002) when analyzing a special LSP, and not the idiolect of a certain person, it is important to include text from different authors. In this corpus, the institutions (universities and legislative organs) are the authors of the texts, so it is important to include text from different universities. In order to gain much information on use of terminology in the area of interest, on one hand, and also to individuate institutional specific terminology, on the other hand, and to assign it to a particular institution, in the building of this corpus it was decided to include texts form all the universities in the countries under investigation.

In the figure below (Fig. 1) the distribution of the number of texts according to the text types and the subdivision of the corpus in a subcorpus with text only from Austria (AT-

corpus), a subcorpus with text only from Germany (DE-corpus) and a subcorpus with text only from Switzerland (CH-corpus) is shown.

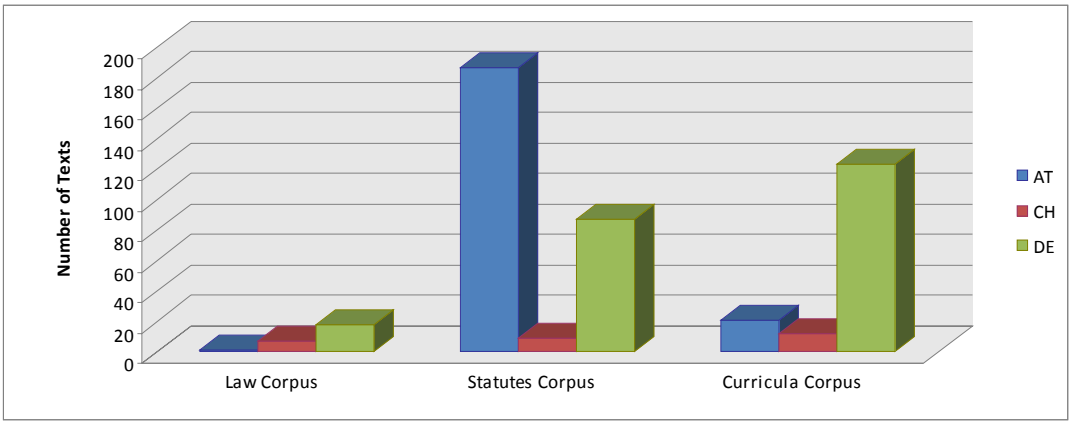

Figure 1. Number of text according to text types and subcorpora.

Due to the different legal situations in the countries under investigation and different numbers of universities, the number of texts is not equal to all of the subcorpora. This fact also has to be taken into account when analyzing and comparing results from the different subcorpora because it influences also the size of the subcorpora as well.

3.2.5 Regional Distribution

Since the UNI-Corpus is not only an LSP corpus or special corpus, but also a special corpus for national varieties of German the regional distribution is an issue. With the help of this corpus variational linguistic research questions in the area of LSP should be studied.

According to Bickel et al. (2009), the localisation of the authorship (for example place of birth, place of death) could be an indicator for the regional distribution of the texts. In the case of the UNI-Corpus it is not possible to individuate single authorship of the normative texts, since different boards and committees are involved in the production of such texts (cf. Hoffmann 1997). However, if the author is not seen as a single person but as an institution,

the authorship can be identified and the texts can be assigned to different universities or legislative organs. That means that the selected texts in the corpus represent the region, canton or federal state or state.

This approach implies that regions, which do not have legislative organs competent in university matters or universities, are not represented in the corpus.

4 Annotations

In corpus linguistics, annotation means assigning linguistic information to the language data of the corpus. Therefore corpus annotation is an added value for a corpus. Leech (1997: 2) states that corpus annotation "is a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development" and it allows the following researchers to base their research on this information and not to start always from scratch. Although different types of corpus annotation, not all annotation types are relevant for the UNI-Corpus. At this stage of the work, only POS tagging and lemmatisation was done. After the conclusion of some researches and studies, other annotation, like annotation of national variants will be included.

POS tagging means assigning a part of speech label to each word in the corpus. This is a process usually done automatically with special tools and programs. The other type of annotation added to the UNI-Corpus is the lemmatisation. Lemmatization is a process that "reduces the inflectional variants of words to their respective lexemes (or lemmas) as they appear in dictionary entries" (McEnery et al 2006: 35). This process usually can be done automatically as well. For the POS tagging and lemmatization of the UNI-Corpus the TreeTagger (Schmid 1994) was used. Since the TreeTagger was developed for German of Germany, some problems while lemmatizing the UNI-Corpus occurred, especially in the Austrian and Swiss subcorpus but also in the German subcorpus when the terms were very

specialized. Similar problems are described in Anstein (2007) with the German variety spoken in South Tyrol.

The statistical TreeTagger could not lemmatize these forms that were missing in his lexicon. Therefore all these words were assigned as "unknown". For this research they might be "potential national variants". Some examples from the "unknown" marked words in the Swiss subcorpus: *Titularprofessor*, *Finanzinspektorat*, *Entlöhnung*, *Legislativkommission*, *Deckungskapitalien*. After checking these words in some dictionaries the following was found: *Entlöhnung*, *Titularprofessor*, *Finanzinspektorat*, *Legislativkommission* and *Deckungskapitalien* were found neither in German dictionary "Wahrig Deutsches Wörterbuch" nor in the Austrian dictionary *Österreichischen Wörterbuch* nor in the variety dictionary *Variantenwörterbuch*. *Entlöhnung* was found in the variety dictionary *Variantenwörterbuch* and it is marked with typical for Switzerland and Luxemburg, and *Inspektorat* was also found and marked as typical for Switzerland and Austria. So there has to be done some further researches but it might be, that *Titularprofessor*, *Finanzinspektorat*, *Legislativkommission* and *Deckungskapitalien* are Swiss variants of specialized terms. All the "unknown" words had to be lemmatized manually.

5. Applications of the UNI-Corpus

With the UNI-Corpus, it is possible to analyse an LSP of one national variety in comparison with another or even two other national varieties of German. The UNI-Corpus opens the possibility to analyse LSP in different national varieties and to find out how much they differ from each other on lexical or on phraseological level.

However, also other research questions that are not focused on national varieties can be pursued with the help of the UNI-Corpus. Internal variation in one of the subcorpora, so only for one variety can be studied starting from a socio(cognitive) terminological approach (cf. Temmerman 2000; Cabré 2003). For example it is possible to look at the variation of the

term *ECTS-Punkt* and its denominational variation in different text types and in the different universities.

6. Conclusions

In this paper it has been shown, that a specialized corpus of national varieties is different from a general reference corpus for national varieties or an LSP corpus and therefore it needs different design criteria. So for example topic is an issue for a LSP corpus and also for a specialized corpus of national varieties. Regional distribution is not a design criterion for general reference corpora or for LSP corpora. In contrast for a specialized corpus of national varieties it is an important criterion because variational linguistic research questions should be analyzed and answered with the help of such a corpus. In this paper it was also shown that a small but well structured corpus can help to study LSP in national varieties and document specialized terms in national varieties. The UNI-Corpus is also suitable for other studies like the study of denominational variation throughout different text types inside one national variety without comparison with another variety.

References

Adamzik, K. (1997). “Fachsprachen als Varietäte”. In: L. Hoffmann et al. (ed.) *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. (Handbücher zur Sprach- und Kommunikationswissenschaft, Bd. 14.1). Berlin: de Gruyter, 522-529.

Aijmer, K. (2008). “Parallel and comparable corpora“. In: A. Lüdeling, (eds.) *Corpus Linguistics. An International Handbook*. Berlin and New York: Walter de Gruyter, 275-292.

Ammon, U. (2004). “Standard Variety/Standardvarietät“. In: U. Ammon (ed.) *Sociolinguistics an international Handbook*. Berlin: de Gruyter, 273-287.

Ammon, U. (1995). *Die Deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Walter de Gruyter: Berlin/New York.

Anstein, S. (2007). "Korpuslinguistische Fallstudien zum Südtiroler Standardschriftdeutsch – das Projekt Korpus Südtirol". Linguistic Online, 32/3, 15-23.

Biber, B., S. Conrad, R. Reppen, (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge Univ. Press.

Bickel, H., M. Gasser, A. Häcki Buhofer, L. Hofer, C. Schön, (2009). "Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten". *Linguistic Online*, 39, 3, 5-31.

Bowker, L. and J. Pearson (2002). *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Cabré Castellví, M. T. (2003). "Theories of terminology: Their description, prescription and explanation". *Terminology* 9:2, 163–199.

Clyne, M (1995). *The German language in a changing Europe*. Cambridge: Cambridge University Press.

Clyne, M. (ed.). 1992. *Pluricentric Languages. Differing Norms in Different Nations*. Berlin/New York: De Gruyter.

Engberg, J. (1993). "Prinzipien einer Typoligisierung juristischer Texte." *Fachsprachen: International Journal of LSP*. 1/2, 31-37.

Hoffmann L. (1997). "Fachtextsorten der Institutionensprachen I: das Gesetz". In: L. Hoffmann et al. (eds.), *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft.* (Handbücher zur Sprach- und Kommunikationswissenschaft, Bd. 14.1), Berlin: de Gruyter, 522-529.

Kjær, A. (1992). "Normbedingte Wortverbindungen in der juristischen Fachsprachen (Deutsch als Fremdsprache)". *Fremdsprache Lerhen und Lernen*, 21, 46-64.

Leech, G. (1997). “Introducing corpus annotation”. In: K. Aijmer and B. Altenberg (eds.) *English Corpus Lingusitics*. London: Longman, 1-18.

Leech, G. (1991). “The state of art in corpus linguistics”. In: J. Svartvik (ed.) *English Corpus Linguistics*, 8-29.

Lemnitzer, L. and H. Zinsmeister (2010). *Korpuslinguistik. Eine Einführung.* Tübingen: Narr Studienbücher.

Markhardt, H. (1993). *Ausdrücke des öffentlichen Bereichs in Österreich*. Brüssel.

McEnery, T., R. Xiao, Y. Tono (2006). *Corpus-Based Language Studies. An Advanced Resource Book*. London/New Zork: Routledge.

Österreichisches Wörterbuch. (2006)[40]. Wien: öbvhpt.

Schmid, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *Proceedings of International Conference on New Methods in Language Processing*. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger (accessed: 10.07.2010)

Sinclair, J. (1996). “EAGLES preliminary recommendations on corpus typology“. Available at: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (accessed: 11.07.2010)

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Temmerman, R. (2000). *Towards New Ways of Terminological Description. The Sociocognitive approach*. Amsterdam/Philadelphia: John Benjamins.

Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. 2004. Herausgegeben von Ammon, Ulrich/Bickel, Hans/Ebner, Jakob/Esterhammer, Ruth/Gasser, Markus/Hofer, Lorenz/Kellermeier-Rehbein,Birte/Löffler, Heinrich/Mangott, Doris/Moser, Hans/Schläpfer, Robert/Schloßmacher, Michael/Schmidlin, Regula/Vallaster, Günter. Berlin: De Gruyter.

Wahrig Deutsches Wörterbuch 2000[7]. Güterloh/München: Bertelsmann.

Wiesmann, E. (2004). *Rechtsübersetzung und Hilfsmittel zur Translation. Wissenschaftliche Grundlagen und computergestützte Umsetzung eines lexikographischen Konzepts*. Tübingen: Günter Narr.