Approaches to Computational Lexicography for German Varieties

Andrea Abel
Stefanie Anstein
European Academy Bozen/Bolzano

Corpora built for linguistic varieties of a pluricentric language such as German are an indispensable resource for a detailed and systematic variety comparison and dictionary development. We present desiderata and suggestions as well as methods from computational linguistics to systematically apply variety corpora for the enrichment, i.e. confirmation, extension and generation, of lexical entries in distinctive variant dictionaries for German. Examples are those variant dictionaries developed by Ammon et al. (2004) and Abfalterer (2007), where we focus on the South Tyrolean German language. On the one hand, we conducted a systematic frequency analysis in newspaper variety corpora for approved lists of South Tyrolean special vocabulary in order to possibly refine corresponding dictionary entries with corpus evidence. On the other hand, we filtered the list of words of our South Tyrolean corpus which could not be lemmatised by a tool developed for the variety in Germany. After removing special vocabulary collected for the South Tyrolean variety in other projects-e.g. legal terms, the remaining list was manually checked for possible new variant dictionary entries, thus-as an innovative variety corpus lexicographic approach-also automatically filtering a huge amount of data to extract only relevant data to be investigated in detail. In addition, we semi-automatically extracted lexical cooccurrences of our two newspaper corpora and compared their frequencies-with the assumption that those cooccurrences are worth being more closely investigated that have high frequency in the South Tyrolean corpus and very low frequency in the corpus from Germany. With these three methods we were not only able to refine dictionary entries for South Tyrolean German, but also to add new ones. The findings on variants can be re-used for further corpus annotation resulting in again better resources for computational variant lexicography of the kind described, which is also to be extended to more complex linguistic levels.

1. Introduction and motivation

More and more corpora are being developed for linguistic varieties of pluricentric languages—and also increasingly for less prominent varieties. Such “variety corpora” as e.g. the *International Corpus of English* (ICE)[1] are an indispensable resource for a detailed and systematic variety comparison as an enhancement of manual studies. For the likewise pluricentric language German (for an overview see Ammon 1995/1997), a corpus initiative of research centres in Germany, Austria, Switzerland, and South Tyrol (Italy) called “C4” is now developing a platform for German variety corpora integrating text data—as balanced as possible—taken from the *DWDS-Korpus*[2], the *Austrian Academy Corpus*[3], the *Schweizer Text Korpus*[4], and the *Korpus Südtirol*[5], the latter of which is the most recent initiative started in the year 2006.

[1] *http://www.ucl.ac.uk/english-usage/ice*; see also bibliography list for specific variety studies

[2] *http://www.dwds.de.*

[3] *http://www.aac.ac.at.*

[4] *http://www.schweizer-texkorpus.ch.*

[5] *http://www.korpus-suedtirol.it.*

Ammon et al. (2004) have published a distinctive variant dictionary for the German language where those lexical items are included which are special to at least one German variety, i.e. they are either exclusively used within one of the German speaking communities, or they share some peculiarities with one or more other varieties. Abfalterer (2007) refined and added entries concerning the South Tyrolean regional variety (see also Egger and Lanthaler 2001) as a part of her PhD thesis and issued a variant dictionary with a special focus on the German variants used in South Tyrol, which play a key role in our preliminary investigations.

The methods used in developing the dictionaries mentioned above consisted basically of manual work—for various reasons such as the lack of digital text corpora and of tools for automated data processing. For new dictionary editions and for modern lexicographic work in general, new possibilities and also requirements on different levels have to be taken into consideration, namely (among others) with regard to (1) content and data modelling, (2) methods for data acquisition and (3) data presentation.

Our investigations are conducted on the basis of related work done within the specific working fields—(1) preceding investigations on the written standard version of the German language in South Tyrol and (2) the domain of corpus lexicography[6] applied for other pluricentric languages. Reacting to desiderata of the former field, our aim is to focus on methodological approaches in computational lexicography and to present preliminary results relevant for lexicographic work. We concentrate on corpus-driven systematic and exemplary case studies on word units and cooccurrences with regard to the South Tyrolean German standard variety and give an outlook for future work in lines with Kilgarriff/Tugwell's (2002) "fourth age of corpus lexicography".

2. Related work

2.1. *Variety corpora*

Work on the varieties of the English language is conducted in the framework of the *ICE*[1] with the aim of enabling researchers to do systematic variety investigations by developing over 20 compatible corpora according to specific corpus compiling criteria.

A corpus specialised on spoken English is the *London-Lund Corpus*[7], for example. The *International Computer Archive of Modern and Medieval English* (ICAME) is an international organisation of linguists and information scientists working with English digitised corpora. Their aim is to collect and provide information on and resources of English language material. Their comprehensive bibliography is available at *http://icame.uib.no/bib_add.html*. A list of more corpus resources for English can also be found at *http://leo.meikai.ac.jp/~tono/resources.html*.

The varieties of French are investigated, e.g., with the *Trésor de la Langue Française Informatisé*[8] *(au Quebec)*—related work is the development of the *Dictionnaire québécois d'aujourd'hui* by Jean-Claude Boulanger.

Spanish in Spain and South America is studied, e.g., with the *Corpus del Español*[9] containing 100 million tokens from the XIIIth to the XXth century coordinated by Mark Davies.

Bacelar do Nascimento et al. (2006) presented specifically related work on a corpora-derived lexicon of Portuguese in Africa.

More detailed and systematic studies are collected, e.g., in Teubert (2001), and extensive bibliographies for comparative studies are presented online (see footnotes 1 or 7). Most of these are,

[6] See also Teubert (2005). Corpus linguistics and lexicography: The beginning of a beautiful friendship.

[7] *http://khnt.hit.uib.no/icame/manuals/londlund/index.htm*; see also bibliography list for specific variety studies

[8] *http://atilf.atilf.fr/tlf.htm.*

[9] *http://www.corpusdelespanol.org*.

however, individual studies without the use of a systematic toolkit[10], the development of which is thus the aim of this research project.

For the German varieties, no such reference data resources or tools have been developed yet - the mentioned initiatives around C4 now provide a promising basis in this respect.

2.2. *Studies and focal points with respect to the German variety in South Tyrol*

As far as research on the written standard (not the dialectal) version of the South Tyrolean German language is concerned, studies have been concentrating on language contact phenomena and particularities on a lexical and also partly on a morpho-syntactical level (e.g. Rizzo-Bauer 1962, Riedmann 1972, Pernstich 1984, Forer and Moser 1988, Lanthaler 1995, Ammon 2001, Ammon et al. 2004, Abfalterer 2007). Extensive investigations on particular features of this variety on the syntagmatic (e.g. collocations, idioms) or the textual level (e.g. Riehl 1997) as well as in-depth examinations of translated texts (e.g. Putzer 1984) are rare, and systematic comprehensive corpus-based studies are still outstanding on all the linguistic levels.

For a long time starting in the 1970s, the interpretation of language contact phenomena has been emphasising on the fact that contact leads to an impairment of the language (e.g. Riedmann 1972). Later on, the focus has been shifted from a research based on the criticism of Italian interferences towards the description of "special vocabularies" from a new, "variety linguistics", point of view. Apart from obvious transfers of words (borrowings and interferences) from the Italian to the German language especially in the fields of public administration and law, South Tyrolean German seems to show—on a purely lexical level—less particularities than assumed (see Pernstich 1984, Ammon 2001).

The single studies on particularities as well as the two existing dictionaries on German varieties already mentioned mainly relied on manual examination and excerption of references (e.g. Riedmann 1972, Pernstich 1984, Riehl 1997). In the cases of the dictionaries of Ammon et al. (2004) and Abfalterer (2007), this was combined with consulting informants as well as with the comparison of data to existing resources such as relevant secondary literature and dictionaries. The more recent studies mentioned were also enhanced to some extent by cross checks in the Internet as a resource for additional evidence (e.g. Abfalterer 2007, Bickel 2000). Finally, recent developments in corpus linguistics, in our specific case the C4 network, presents an indispensable basis both for extensive research on German varieties and for concrete lexicographic work.

2.3. *Requirements for variety corpus lexicography*

In the following, some research desiderata as well as prerequisites and requirements in the field of lexicography for German varieties will be presented, as well as concrete recommendations for practical lexicographic work on German varieties.

As can be deduced from the actual state of research on South Tyrolean German, large-scale investigations on this standard variety 1) are necessary on a lexical and especially also on a syntagmatic and on a textual level, 2) by intralinguistically comparing it to other varieties, and 3) using state of the art corpus linguistic methods and technologies.

In addition, based on the exemplary use of these technologies for the South Tyrolean written standard German, we can formulate general desiderata for lexicography dealing with language varieties on different linguistic levels.

Three requirements that have to be taken into consideration for corpus lexicography are described in the following.

1) With regard to content, corpus data can help to confirm and/or supplement and enrich existing lexical entries. Furthermore, new entries can be added as huge corpus data

[10] ICECUP (ICE Corpus Utility Program) is an example for an automated, but not primarily comparative tool: *http://www.ucl.ac.uk/english-usage/resources/icecup/*.

allows us to detect new lemma "candidates" and to get more evidence on specific lexical items and their peculiarities. While in this paper we mainly focus on these aspects, in a next step an extension of the dictionaries' data models could be envisaged. This could be the integration of new data categories, e.g. the extension of pronunciation data, the insertion of collocations/coocurrences and idioms including the addition of lexicographic examples for each of them, further special notes and comments (on specific uses of variants), frequency labels, or diachronic indications as a documentation of language change. Most of this is data which obviously can only be integrated in a dictionary on the basis of huge, balanced corpora whose data are evenly spread over text types and decades within a certain time period.

2) A further important aspect are the concrete methods for data acquisition applying computational linguistic methods. The improvement and refining of existing tools as well as the development of new specific tools make it possible to compare varieties and variants in a semi-automatic way and thus facilitate and support the manual work that of course still has to be done by the lexicographer. In section 3, three of our approaches are presented in some detail.

3) In addition, data presentation has to be taken into consideration as an important issue. When dealing with printed dictionaries, a clear discrimination of data categories by using colours etc. should be aimed at. With respect to possible online versions, direct links to corpus data (with lemmata or cooccurrences as a starting point) is e.g. an interesting feature for the user searching for further evidence autonomously.

Taking all this into account in order to further enrich the existing dictionaries and work on new editions with the help of corpus resources and analysis tools from computational linguistics, we are developing methods to semi-systematically apply variety corpora for the confirmation, extension and generation of lexical entries for the South Tyrolean German language.

3. Methods

3.1. *Data*

We used—as our two text corpus resources for exemplary comparative variety studies—the South Tyrolean daily newspaper *Dolomiten* (abbreviated: Dolo; between 1991 and 2006) containing approximately 66 million tokens and the German newspaper *Frankfurter Rundschau* (abbreviated: FR) from 1992 to 1993 with about 40 million tokens. Both corpora were tokenized, part of speech tagged, lemmatized and chunked using TreeTagger (Schmid 1994) and YAC (Kermes 2003). The corpus query system used is the Corpus Query Processor CQP (Christ 1994, Evert 2005). Further resources are the findings of the project *Datenbank zum Südtiroler Deutsch*[11] (2004) and special word lists from the variant dictionaries Ammon et al. (2004) and Abfalterer (2007: *Südtirolismen*), as well as lists with special vocabulary such as names or South Tyrolean legal terminology.

3.2. *Südtirolismen*

In a first approach, starting from lists of so-called *Südtirolismen* (*South Tyrolisms*: special vocabulary of South Tyrol; Abfalterer 2007), we compare the frequencies of these words in the corresponding variety corpora to each other. We investigated those *South Tyrolisms* which appear with unexpectedly high frequency in the corpus from Germany in more detail, as we assumed a possible refinement of the variant dictionary entries for these cases.

[11] Database for South Tyrolean German; *http://www.uibk.ac.at/projects/woerterbuch/sued/sued.html*.

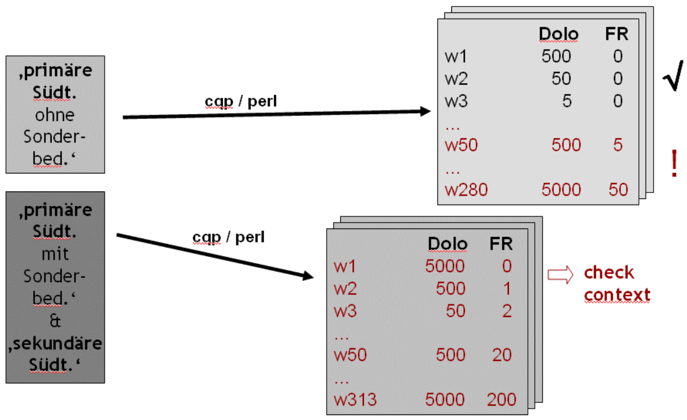

Figure 1. The “primary South Tyrolisms without special meaning” which are supposed to occur only in South Tyrol (above) are counted and compared in the two corpora. If the FR holds occurrences, these are investigated in detail, as well as the “primary South Tyrolisms with special meaning”, which are only in a special meaning real South Tyrolisms, and which together with the “secondary South Tyrolisms” were studied by also looking at their sentence context.

We conducted the variety corpus comparison depicted in figure 1 using the corpus query system CQP combined with Perl scripts for the systematic processing of query results. Table 1 shows a resulting extract of an example output list for “primary South Tyrolisms without special meaning” sorted according to frequencies in FR.

	Dolo	vs.	FR
Abgeordnetenkammer	801		57
Schulamtsleiter	836		48
SVP	14566		29
Regionalregierung	504		25
Proporz	754		21
Regionalrat	382		8
Finanzpolizei	519		7
Lyzeum	205		6
Vertrauensarzt	59		5
Industriellenverband	587		4
Schriftleiter	42		3
Sanitätseinheit	394		3
Carabiniere	349		3

Table 1. Frequencies for “primary South Tyrolisms without special meaning”

Our corpora are obviously neither balanced nor big enough to draw conclusions for general language usage. We have not used statistical measures yet, but for first studies we work with absolute frequencies, which give valuable hints and automatically constrict the set of data that has to be evaluated manually. A report on the results will be given in section 4.

3.3. Unknowns

In our second substudy, we filtered the list of words of our South Tyrolean newspaper corpus which could not be lemmatised by the simple version of the TreeTagger. As this tool and its lexicon was designed for and trained on the variety in Germany, the assumption was that new candidates for South Tyrolean variants could be found in that list of *unknowns* (see figure 2). As filters, we used existing word lists with German, Austrian, Swiss and South Tyrolean special vocabulary, terminology of the bistro database[12], proper names, and foreign word dictionaries to reduce the list of *unknowns*.[13] The remaining list was then manually checked for lemmata to create possible new dictionary entries or to be added to the filter lists, which can be used again for corpus annotation or, e.g., spell-checking. On the results see section 4.

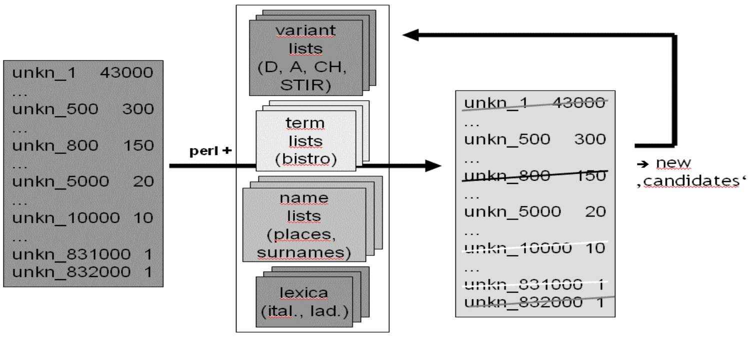

Figure 2. Filtering of the *unknowns* in the Dolomiten corpus yielding new "candidates"

3.4. Direct *and* indirect *cooccurrences*

On a more complex linguistic level, for the analysis of recurrent linguistic patterns, we extracted cooccurrences from the two corpora again with the help of CQP and Perl as shown in figure 3. We did this on two levels—one were *direct* cooccurrences such as Adj+N or Prep+N, which we investigated systematically, and the others were *indirect* cooccurrences with syntactic relations such as Subject/Object+V, where we were able to do exemplary studies. For the latter, several scripts extracted patterns from different sentence structures and combined them to a single frequency list for each pattern and corpus. These frequency lists of the two corpora were then confronted with each other, again with the assumption that South Tyrolean particularities do not or very rarely occur in the FR corpus and that it is worth looking at these cases in particular.

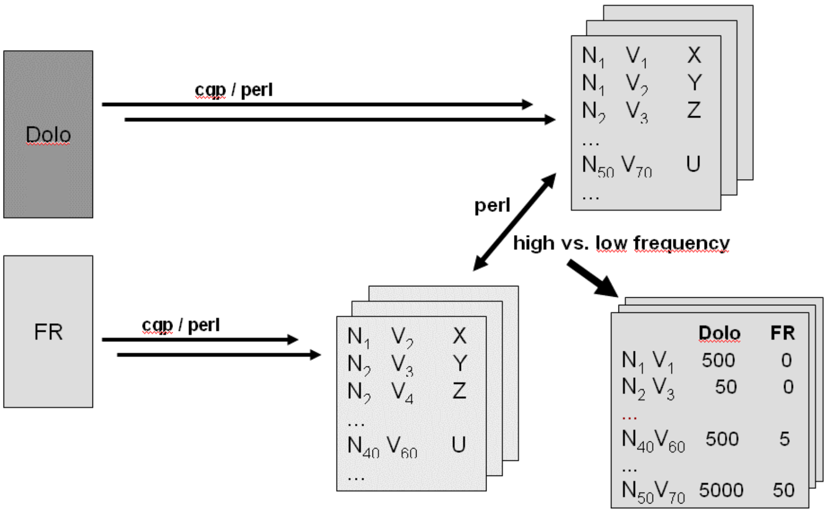

Figure 3. Extraction and comparison of cooccurrences in the two corpora

[12] *http://www.eurac.edu/bistro*.

[13] Thanks to Alexandra Hagelstein and Stefanos Petrakis for their support of this work.

On the one hand, collocations that differ with respect to the German varieties can be yielded (e.g. noun verb collocations), and on the other hand, special cooccurrences for *South Tyrolisms* can be found to add them to the *South Tyrolism* dictionary entries. Some results of these studies are presented in the following section.

4. Results and examples

Our method of automatically filtering a huge amount of data in order to extract only relevant data to be investigated manually provided the following first results:

For e.g. the South Tyrolism *Abgeordnetenkammer* we can confirm its usage by various further result sentences from the corpus, which can also be included in the dictionary.

We can secondly enrich dictionary entries and add information (e.g. special senses of words for the South Tyrolean variety), as it is the case for e.g. *Konsortium* ("consortium"), where we suggest a less specific definition than the mentioned variant dictionaries contain at present, namely in this case closely following the definition of the Italian corresponding word.

More examples are the adverb *weiters* and the preposition *ober*, both *Austriacisms*, which are also used in South Tyrol and should get this label in a new dictionary edition. This was confirmed by their high frequency in the Dolomiten corpus (2778 and 231, respectively).

In addition, we can add new dictionary entries for South Tyrolean German, e.g. for the adverb *ehestens* ("as soon as possible" with frequency 123).

For the last study on cooccurrences, we found various adjective noun combinations in the Dolomiten corpus, such as *allgemeine Klasse* ("general class") for sports or *weißer Stimmzettel* ("white/empty ballot"), which do not occur in the FR corpus and are thus candidates for collocation entries. For combinations with prepositions, we noticed e.g. the interference from Italian using the local preposition *innerhalb* ("within") in a temporal context (*innerhalb Januar* "within January"), a usage where the degree of "correctness" is to be discussed.

In an additional case study on collocators of South Tyrolisms as bases, we identified several typical collocators for the noun *Mobilität* (which has the particular sense "dismissal/unemployment" only within the South Tyrolean variety), which are e.g. "jemanden in die Mobilität *überstellen*" ("to *transfer* someone into unemployment") or "sich *in* Mobilität *befinden*" ("to be unemployed").

For functional cooccurrences in general such as subject and object combinations, single examples could be interpreted—after that we assume not too much variation in general, which is however to be investigated with more fine-grained tools (e.g. Heid/Ritz 2005) and more manual work.

5. Summary and outlook

With the presented semi-automatic approach, more data can be investigated than could be done purely manually, while it is obvious that manual work is still necessary for detailed semantic interpretation and thorough lexicographic work. We were able to confirm as well as enrich existing lexicon entries and make suggestions for adding new ones.

The corpora taken as a basis have to be enhanced and to be made as comparable as possible, which is one aim of the C4 initiative mentioned. We will also concentrate on developing more tools for the semi-automatic comparison of varieties on the basis of corpora along the lines of the approaches presented in this paper. Existing exemplary findings will be systematized to be used as a further basis for investigations. The quality of such automated comparison results depends also substantially on the quality of the linguistic annotation, which is in some cases a challenge for tools that are created for the more frequently used varieties in contrast to the lesser-used varieties. So more research and development of semi-automatic tools is also to be done there.

The quantitative and qualitative investigation of phenomena will e.g. include *Südtirolismen* and their collocators, still unclear entries in the *Datenbank zum Südtiroler Deutsch*, phraseologisms and their degree of fixedness and idiomaticity, the comparison of synthetical vs. analytical constructions (e.g. compound *…-ministerium* vs. *Ministerium für…*), differences in pragmatics, discourse, etc., and also investigations on “cause” and “origin” for certain phenomena such as language contact or language variation over time.

In a further step, all the findings on variants can again be used to enhance and improve the annotation of variety corpora and to adapt annotation tools. This results in still better outcomes of further corpus analysis and comparison studies, which are the basis for the presented methods for a valuable systematic enrichment of variant dictionaries. These approaches to computational variant lexicography can then also successfully be applied to more complex linguistic levels to extract collocations or to work in the field of phraseology, important material also to be included in future variant dictionaries.

Bibliography

Abfalterer, H. (2007). Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols. Innsbruck: Innsbruck University Press.

Ammon, U. (1995). Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten. Berlin: de Gruyter.

Ammon, U. (1997). *Nationale Varietäten des Deutschen*. Heidelberg: Julius Groos.

Ammon, U. et al. (2004). Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin: Walter de Gruyter.

Bacelar do Nascimento, M. F. et al. (2006). "The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-Derived Lexicon". In Calzolari, N. et al. (eds.). *Proceedings of the 5th International LREC (Genoa, Italy)*. 1791-1794.

Bickel, H. (2000). "Das Internet als Quelle für die Variationslinguistik". In Häcki Buhofer, A. (ed.). *Vom Umgang mit sprachlicher Variation: Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte*. Tübingen – Basel: Basler Studien zur deutschen Sprache und Literatur 80. 111-124.

Christ, O. (1994). "A Modular and Flexible Architecture for an Integrated Corpus Query System". In Proceedings of COMPLEX 1994, 3rd Conference on Computational Lexicography and Text Research Budapest, Hungary, July 7-10. 23-32.

Egger, K.; Lanthaler, F. (eds.) (2001). Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt. Wien: Folio.

Evert, S. (2005). The CQP query language tutorial. Technical report. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. *http://www.ims.uni-stuttgart.de/projekte/ CorpusWorkbench*.

Forer, R.; Moser, H. (1988). "Beobachtungen zum westösterreichischen Sonderwortschatz". In Wiesinger, P. (ed.). *Das österreichische Deutsch*. Wien – Köln – Graz: 189-209.

Heid, U. (2001). "Collocations in Sublanguage Texts: Extraction from Corpora". In Wrigth, S. E.; Budin, G. (eds.). *Handbook of Terminology Management*. Amsterdam: John Benjamins. 788-808.

Heid, U.; Ritz, J. (2005). "Extracting collocations and their contexts from corpora". In Pajzs, J. et al. (eds.). *COMPLEX 2005*. Budapest: Linguistics Institute, Hungarian Academy of Sciences. 107-121.

Kermes, H. (2003). *Off-line (and On-line) Text Analysis for Computational Lexicography*. Dissertationsschrift. Stuttgart: Universität Stuttgart.

Kilgarriff, A.; Tugwell, D. (2002). "Sketching words. Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. Marie-Hélène Corréard (Ed.) EURALEX: 125-137.

Lanthaler, F.; Saxalber, A. (1995). Die deutsche Standardsprache in Südtirol. In Muhr, R.; Schrodt, R. et al. (eds.). Österreichisches Deutsch. Linguistische, sozialpsychologische und sprachliche Aspekte einer nationalen Variante des Deutschen. 289-305.

Lanthaler, F. (2001). "Zwischenregister der deutschen Sprache in Südtirol". In Egger, K.; Lanthaler, F. (eds.). *Die deutsche Sprache in Südtirol: Einheitssprache oder regionale Vielfalt*. Wien: Folio. 137-152.

Pernstich, K. (1982). "Deutsch-italienische Interferenzen in der Südtiroler Presse". In Moser, H. (eds.). *Zur Situation des Deutschen in Südtirol. Sprachwissenschaftliche Beiträge zu den Fragen von Sprachnorm und Sprachkontakt*. Innsbruck. 91-182.

Putzer, O. (1984). Interferenz in Übersetzungen: Aspekte der Übersetzungsleistungen bei der Zweisprachigkeitsprüfung (D.P.R. 752/76). Bozen: Assessorat für Schule und Kultur in italienischer Sprache.

Riedmann, G. (1972). Die Besonderheiten der deutschen Sprache in Südtirol. Mannheim.

Rizzo-Bauer, H. (1962). Die Besonderheiten der deutschen Schriftsprache in Österreich und Südtirol. Mannheim.

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

Teubert, W. (ed.) (2001). Text Corpora and Multilingual Lexicography. Special issue of International Journal of Corpus Linguistics.

Teubert, W. (2005). “Corpus linguistics and lexicography: The beginning of a beautiful friendship”. In Wiegand, H. E. et al. (eds.). *Lexicographica: Internationales Jahrbuch für Lexikographie 2004*. Tübingen: Niemeyer.