Comparing Geographical and Learner Varieties on the Basis of Corpora

Stefanie Anstein Aivars Glaznieks

Institute for Specialised Communication and Multilingualism
European Academy Bolzano / Bozen
stefanie.anstein@eurac.edu, aivars.glaznieks@eurac.edu

Résumé Dans cet article, nous présentons des études systématiques de deux types de variétés de langage pour leur comparaison et documentation. Pour analyser des variétés géographiques de l'allemand, nous avons employé le *Korpus Südtirol* annoté dans le cadre du *C4*. L'outil de comparaison *Vis-À-Vis* extrait semi automatiquement des particularités propres aux variétés, en alliant méthodes quantitatives et qualitatives afin de faciliter et de réduire le travail manuel des linguistes. Dans le cadre du projet relié *KoKo*, des corpus de textes écrits par des apprenants germanophones de trois différentes régions d'Italie, d'Autriche et d'Allemagne sont comparés. Les analyses se concentrent sur l'emploi des différentes variétés d'allemand dans un cadre éducatif ainsi que sur les facteurs de détermination linguistiques et sociolinguistiques qui influencent les compétences d'expression écrite des étudiants. Nous décrivons nos méthodes et outils de linguistique de corpus de même que quelques premiers résultats à différents niveaux de la description linguistique.

Abstract In this article, we present systematic studies of two kinds of language varieties for their comparison and documentation. To analyse geographical varieties of German, the annotated *Korpus Südtirol* in the framework of the *C4* initiative is used. The comparison toolkit *Vis-À-Vis* semi-automatically extracts varieties' particularities in order to support and reduce linguists' manual work, combining quantitative methods with qualitative ones. In the related project *KoKo*, written text corpora of German-speaking learners of three different areas in Italy, Austria, and Germany are compared. The analyses focus on the use of different varieties of German in an educational setting as well as on determining linguistic and sociolinguistic factors influencing the students' writing competences. We describe our corpus linguistic methods and tools as well as some first results on different levels of linguistic description.

Mots-clés : variétés géographiques de langage, corpus d'apprenant, comparaison de variétés, linguistique computationnelle, linguistique de corpus

Keywords: geographical language varieties, learner corpora, comparison of varieties, computational linguistics, corpus linguistics

1 Introduction and background

Varieties of a language, either geography- or proficiency-related, have many similar characteristics, which is why it is crucial to extract subtle differences between them. These are relevant e.g. for variant lexicography (cf. Nelson, 2006) or for evaluating language proficiency in both L1 (cf. Schneider, 2001) and L2 (cf. Spiekermann, 2007) to support language teaching and learning (cf. Granger et al., 2002). The variety we focus on is used in the Autonomous Province of Bolzano / South Tyrol in Northern Italy, which is a bilingual German-Italian region where a variety of the pluri-centric language German (cf. Ammon, 2005) is prevalent.

The general aims of this article are to describe two major projects on the comparison of varieties on the basis of corpora, to describe and reflect the methods used and the tools developed, as well as to present first results of the comparisons. In this inherently cross-disciplinary field, we show how methods of computational linguistics, sociolinguistics, and contrastive linguistics can be combined.

1.1 The language situation in South Tyrol

In South Tyrol, two thirds of the around 500.000 inhabitants are mother-tongue German speakers, most of them using local dialects. Regulated by the Second 'Autonomiestatut' in 1972 (esp. Art. 99, 100; cf. Autonome Provinz Bozen-Südtirol, 2006), German is an official language besides Italian (and in some valleys Ladin), so they are equally used in all areas of public life - in the media, in the educational system, and in public administration -, which results in multilingual publications of all kinds, from laws and official letters up to street signs. The South Tyrolean dialects are predominantly used by the German-speaking population in both oral and written interpersonal communication. With consequence of a diglossic situation, the standard variety is constricted to the educational system (during class), to communication with foreigners, and to written texts that appear in public.

The South Tyrolean written standard language shows several differences compared to the other varieties of German, partly caused by the South Tyrolean dialect, by the contact language Italian, and by the neighbouring variety of Austrian German. Phenomena on the lexical level are often a result of the political situation of South Tyrol being part of Italy, as it was necessary to loan and translate Italian expressions or to extend meanings of German words in order to transfer Italian law and administrative terms. The particularities of the South Tyrolean variety on other levels of linguistic description has not been comprehensively investigated (see section 2.1); unique aspects on the morphological and syntactic level, for example, were often considered as individual mistakes.

1.2 Corpus linguistic projects and resources

Corpus linguistics is a branch of Natural Language Processing that deals with the collection and analysis of large amounts of authentic language data (cf. Lüdeling, Kytö, 2009).

In the long-term initiative *Korpus Südtirol,* we are collecting and analysing written German in South Tyrol (cf. Abel, Anstein, Petrakis, 2009). The current corpus (queriable at http://www.korpus-suedtirol.it) comprises several linguistically annotated corpora of around 70 million tokens. *Korpus Südtirol* is a highly relevant documentation of the written language

in South Tyrol contributing also to the historical and cultural heritage of the region. The main aim of the initiative is to provide an empirical basis for the debate on the linguistic situation in South Tyrol. To supplement, improve and revise earlier manual investigations, frequency-based and statistical analyses of differences between various corpora can be conducted. The next step of such an analysis is to interpret the results qualitatively within a linguistic and sociolinguistic framework and investigate causes and consequences.

One aim of the recently launched project *KoKo* is to analyse and compare the writing skills of learners from such a perspective. We collect essays written by learners in the last period of their academic education. Since the standard variety in South Tyrol is used in written language in all official contexts, it is crucial for pupils and students to master the standard variety, particularly with regard to the labour market. In *KoKo,* students from all school types that qualify for university entry are included. The essays will constitute an extensible learner corpus and contribute thus both to the documentation of the German language in South Tyrol and to its investigation. Besides in South Tyrol as the focus of our interest, learner texts will also be collected in the state of Tyrol (Austria) and the Free State of Thuringia (Germany) for reasons of comparison. In addition, the corpus will provide metadata to analyse the influence of extra-linguistic variables on the participants' language skills.

2 Related work

Text corpora serve as a valuable basis to semi-automatically identify relevant particularities of varieties and differences among them. Below, we report on some earlier findings on South Tyrolean German as well as present several studies in variety and learner corpus linguistics.

2.1 South Tyrolean German

Single studies of the South Tyrolean German have been conducted since the 1960s (e.g. Rizzo-Baur 1962, Riedmann 1972), focusing rather critically on interferences of Italian. Later, this attitude shifted towards a less judgmental interpretation and to the description of official variants. The first comprehensive variant dictionary for German, the *Variantenwörterbuch* (Ammon et al. 2004), and its extension for South Tyrolean German (Abfalterer, 2007, which comprises an extensive list of South Tyrolean variants used in all areas of everyday life), were elaborated with the help of many experts cross-checking texts from all the varieties. Studies towards automating such investigations with the system *Vis-À-Vis* started in 2007; the latest reports are Anstein (2009) and Abel, Anstein (2010).

There are many examples for the particularities of the German standard variety in South Tyrol, and all levels of linguistic description are affected. At the lexical level, differences with respect to the other German varieties usually consist in one-to-one equivalents such as *Kondominium* ('apartment building'; cf. it. *condominio*) in South Tyrol vs. *Mehrfamilienhaus* in Germany, or in many-to-one equivalents such as *provisorische Ausfahrt* ('temporary exit'; cf. it. *uscita provvisoria*) vs. *Behelfsausfahrt,* respectively. More complex phenomena such as differing collocations (e.g. *weißer Stimmzettel* vs. *ungültiger Stimmzettel* 'void ballot'; cf. it. *scheda bianca*) or subtle semantic differences[1] up to pragmatic particularities of a variety are more difficult to detect and extract. Phenomena on the morpho-syntactic level have also been

[1] The word *Mobilität*, e.g., which means 'mobility' in general, has extended its meaning in South Tyrol and refers additionally to a special kind of unemployment.

described; especially differences between dialect and standard variety with respect to the inflection of articles and nouns has been mentioned by several researchers (e.g. Egger, 1979: 72-78; Giacomozzi, 1982: 79-84). For example, in the dialect there is no distinction between accusative singular (*in Bua* ('the boy'$_{acc}$)) and dative singular (*in Bua* ('the boy'$_{dat}$)) forms of the definite article of masculine nouns, whereas the standard makes a distinction (*den Bub(en)* vs. *dem Bub(en)*; cf. Giacomozzi, 1982: 90-97). This difference between dialect and standard often causes an unsureness in the use of accusative and dative forms that can be observed even within prepositional phrases (e.g. **Interesse and den*$_{acc}$ (instead of *dem*$_{dat}$) *Service* 'interest in the service'). Other phenomena that are affected are differences in gender or the inflection of verbs. In addition, syntactic particularities are discussed to be a result of the language contact to Italian. They concern word order, the position of the inflected verb, the conjunction of sentences, and constructions with the participle (cf. Egger, 1979: 84-97).

2.2 Variety corpus linguistics

Comparable corpora for geographical varieties are being compiled in projects such as the well-known *International Corpus of English (ICE)*, the *Trésor de la Langue Française Informatisé (au Quebec)*, or the *Corpus del Español*. Also for German, an initiative of research centres in Austria, Germany, South Tyrol, and Switzerland called *C4*[2] is developing variety corpora which are comparable with respect to contents and size.

Related work has furthermore been done in diachronic linguistics on the comparison of language over time (cf. Janda and Joseph, 2004), of originals and translations (cf. Baroni and Bernardini, 2006), or of native and second language (cf. Netzel et al., 2003), to name but a few. Especially many of the earlier studies were conducted manually and often for very specific phenomena. Our aim is now to provide a toolkit that supports the systematic and comprehensive comparison of language varieties.

2.3 Learner corpus linguistics

One of the L1 learner corpora for written German is the so-called *Ludwigsburger Aufsatzkorpus* which was provided by Fix, Melenk (2000). It consists of 2300 essays that have been used in some recent investigations (e.g. Margewitsch, 2006). However, there are only a few learners' databases that consider L1, mostly for spoken language, e.g. CHILDES (*Child Language Data Exchange System*). There are more databases focusing on L2, e.g. FALKO (*Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*), VALICO (*Varietà die Apprendimento della Lingua Italiana: Corpus Online*), ICLE (*International Corpus of Learner English*), and FRIDA (*French Interlanguage Database*).

Augst, Faigel (1986), e.g., published a study in which they investigated the development of writing skills in L1 of 13-23-year-olds. The participants were asked to express their position on a given topic; in addition, oral statements to the same topic were required by a small number of the participants. All collected texts were analysed on the lexical and syntactic linguistic level as well as on the structural and compositional level of the texts. Results show a clear enhancement with age in the distinction between oral and written language with respect to the analysed linguistic levels. A very useful summary of related work from an ontogenetic perspective can be found in Becker-Mrotzek, Bötcher (2006).

[2] http://www.korpus-c4.org

There are several studies that have analysed essays written by German-speaking pupils and students from different perspectives (e.g. Sieber, 1998). Recently, Dürscheid, Wagner, Brommer (2010) have published an investigation in which they analyse the influence of frequent writing in extracurricular activities (e.g. writing short messages and e-mail, participating in social networks and chats) on writing skills in school. One of their research questions was whether or not the dialect that is frequently used in extracurricular written communication among Swiss learners influences the writing skills in the standard variety. Results show that the essays were frequently interspersed with dialectal expressions, although participants clearly separate extracurricular from curricular writing with respect to style and typographical particularities.

Existing analyses suggest that there are many factors influencing the way students write. In recent years, extra-linguistic variables were often included in data collection (e.g. Abel, Vettori, Wisniewski, in preparation). However, automatic analyses of large corpora often were not possible while manual analyses of small corpora were preferred for investigating learners' writing skills. This stresses the timeliness of the projects presented in this paper: Comprehensive, systematic, and semi-automatic investigations with a special consideration of sociolinguistic factors of the South Tyrolean region are definitely needed.

3 Methodology

In this chapter, we briefly describe the resources, methods, and tools used for the comparison of language varieties in the two projects described.

3.1 Corpus design

Korpus Südtirol contains an extensive collection of texts written in German by South Tyrolean authors. In accordance with the *C4* project, it comprises texts of four types: fictional and non-fictional texts, functional texts, and journalistic prose. In addition, comprehensive metadata on the author and the publication was collected. Currently, the corpus consists of almost 300.000 texts and around 70 million tokens and is continuously being increased (cf. Abel, Anstein, Petrakis, 2009).

For *KoKo*, a text production task in class and a sociolinguistic survey will be conducted. We aim at collecting 600 learner essays from the region of South Tyrol and additional 600 essays from Tyrol and Thuringia, respectively. The essays will be written during class as part of the curriculum and will be graded by the teachers. For the sake of comparison, we will determine the topic of the essay, but the situation remains an authentic test at school. The metadata consist of biographical and socio-economical information of the participants and their parents. In addition, participants' habits with respect to language production and perception will be recorded. In particular, we focus on the usage of various registers of spoken German (dialect vs. colloquial vs. standard variety). Finally, we are interested in the participants' evaluations of the usages of dialect and standard in formal and informal situations. We assume that this data will show which variables influence the participants' written language.

All the corpora are pre-processed with standard annotation tools on the word level (e.g. Tree Tagger; Schmid, 1994) and prepared to be queried and statistically analysed with the Corpus Query Processor (CQP; Christ, 1994).

3.2 Comparison approach

For a systematic and comprehensive comparison of corpora on different levels of linguistic description, semi-automatic tools are needed, since manual evaluation is time-consuming and costly. Resources such as corpora have to be compared directly according to regular patterns with statistical counts, and on different levels of linguistic description, where the comparability of the contents and of the corpora in general has to be taken into account (cf. Gries, 2007). Automatic filtering of statistically produced lists containing suggested 'candidates' for differences or particularities reduces manual work and supports experts in their evaluation, who can then concentrate on the interpretation of the remaining, mostly new phenomena.

The analyses of the students' essays require standardized and comprehensible criteria. The *Zürcher Textanalyseraster* ('Zurich analyzing pattern for texts'; cf. Nussbaumer, Sieber, 1994) for example, a very systematic and objective way of analyzing language skills, considers quantitative and qualitative aspects. It reflects the authors' conception of an accurate text and considers formal mistakes, stylistic aspects, as well as the functional appropriateness of the text. Since all linguistic levels of description are scheduled for the analysis, from orthographical correctness and morphological as well as syntactical accurateness to the analysis of the semantics of simple and complex expressions, the pattern constitutes a perfect basis to develop an adjusted pattern for the comparison of learner varieties.

3.3 The system *Vis-À-Vis*

Our toolkit for the systematic comparison of varieties on the basis of corpora consists of several independent modules (the exact procedures of it cannot be described here, for details see Anstein, 2009). In Figure 1, the workflow of the system and its functionalities are shown. In the central modules, the corpora are analysed and compared with a combination of existing as well as new or adapted tools, symbolic as well as statistic ones. First, lexical frequency statistics are applied. Further modules are elaborating on morphology, collocations, phrases, or syntactic features, up to more subtle semantic differences, textual features, or pragmatics.

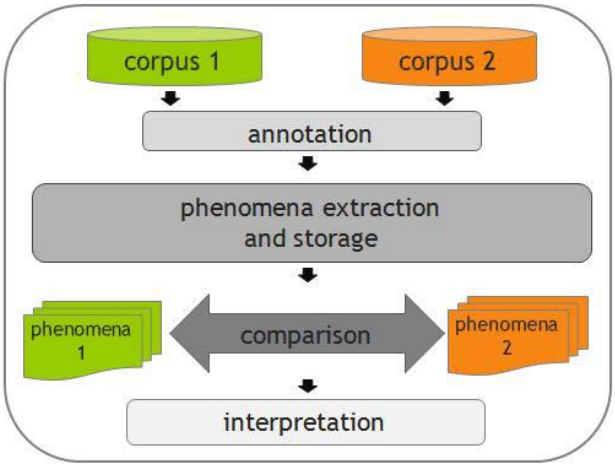

Figure 1: Overall architecture of *Vis-À-Vis*

As input, users provide the corpora to be compared, which are then annotated with standard tools. Here, difficult cases for the tools or errors produced can identify the first set of

candidates for special variety characteristics, since the tools are usually created for the 'main' varieties. As a result, *Vis-À-Vis* produces filtered lists of probably relevant differences between the varieties on different levels of linguistic description for manual evaluation, which is necessary for the verification of phenomena. Trivial characteristics of a variety or knowledge already investigated and confirmed (e.g. collections of proper names or regionalisms) are automatically removed from candidate lists. The data will be presented in a user-friendly and intuitive way to facilitate their interpretation and further processing for variety and learner corpus linguistics; for sentence contexts of ambiguous or other difficult cases, it will be possible to search directly in the annotated corpora. In a further step, the findings such as lexical difference lists or striking syntactic patterns can again be used for the annotation of special phenomena in other corpora.

4 Preliminary results

If we contrast frequencies in a South Tyrolean and a German newspaper[3] corpus for example, lists of adjective noun cooccurrences with high frequency in South Tyrol and lower frequency in Germany can be produced. An extract of such a list is shown in table 1. Since it is not yet filtered manually, it shows both linguistically relevant differences as well as differences that might reflect the selection of the topics in the respective corpora or situational specificities, e.g. an economic focus of a region the interpretation and more detailed corpus analysis based on these results can then be conducted by linguists. Other corpus-based findings on South Tyrolean particularities with a detailed discussion are described in Abel, Anstein (2010).

	ADJ + N	**absolute frequency in DOLO[5]**	**relative frequency in DOLO[5]**	**absolute frequency in FR[5]**	**relative frequency in FR[5]**
1	**grünes Licht**	988	1.4891	314	0.7811
2	**vergangene Saison**	985	1.4846	280	0.6965
3	**freiwilliger Helfer**	976	1.4710	134	0.3333
4	**ganze Welt**	963	1.4514	265	0.6592
5	**kommende Saison**	959	1.4454	259	0.6443
6	**zweiter Durchgang**	951	1.4333	188	0.4676
7	**zweiter Teil**	912	1.3746	223	0.5547
8	**öffentliches Verkehrsmittel**	892	1.3444	175	0.4353
9	**erster Durchgang**	890	1.3414	131	0.3259
10	**morgiger Donnerstag**	868	1.3083	87	0.2164
11	**erster Lauf**	866	1.3052	58	0.1443
12	**erster Sieg**	856	1.2902	163	0.4055
13	**heuriges Jahr**	839	1.2645	0	0

Table 1: adjective noun cooccurrences contrasted between variety corpora

[3] DOLO = ca. 66 million tokens of the South Tyrolean daily newspaper *Dolomiten*;
FR = ca. 40 million tokens of the German daily newspaper *Frankfurter Rundschau;*
*relative frequency = absolute frequency / corpus size * 100.000*

With respect to learner corpora, preliminary analyses of essays collected in the related project *KOLIPSI L1* point to promising findings. In addition to the unsureness regarding case markers described above, the particularities include syntactic constructions and formulaic language as well as idiomatic expressions and specific characteristics on the lexical level in general. With respect to syntactic constructions, one particularity was the use of embedded bare-infinitival interrogatives. According to Sabel (2006: 245), the construction is not productive in the standard variety in Germany and restricted to the form *Ich weiß nicht was tun* 'I don't know what to do' (cf. also Reis, 2003: 173-174). In our preliminary investigation, however, we found various instances of this construction. In contrast to the variety in Germany, it is productively used in spoken German in South Tyrol but has not been described for the standard variety. Since it is a very productive construction in Romance languages like Italian (e.g. *non so dove andare* 'I don't know where to go'), interference might be a possible cause. Such constructions are not expected to be found in the texts written by students from Austria and Germany. However, a systematic comparison is crucial to evaluate and interpret the findings and to differentiate between common features and regional particularities of the German language.

5 Conclusion

This paper briefly describes ongoing work in two related projects to compare language varieties on the basis of corpora, where manual expert comparative work is supported semi-automatically. With the toolkit *Vis-À-Vis*, systematic differences between varieties can be detected and verified. This is to base the standard and norm discussion on empirical data, to support language didactics with adaptable material, and especially in our context to raise general awareness for South Tyrolean particularities.

For the analyses of lerners' texts within the project *KoKo,* an analyzing pattern similar to the *Zürcher Textanalyseraster* will be developed to guarantee an objective qualitative evaluation. The work with the pattern will benefit from the semi-automatic comparison of the corpora with *Vis-À-Vis.* The results of the project are crucial for a better understanding of learners' problems with the standard variety in South Tyrol. Based on our findings, we will be able to make specific suggestions for adapted language didactics and thus counteract the unsureness towards the standard variety of many South Tyroleans. In addition, the project also helps to promote the language awareness of the native speakers of German in South Tyrol, which in turn seems to be crucial in both the diglossic and the bilingual environment of the region.

Acknowledgements

Thanks to all contributors to *Korpus Südtirol* and *KoKo,* esp. the head of the projects Dr. Andrea Abel, the European Academy Bolzano / Bozen, the Free University of Bolzano, the University of Innsbruck, and the Ministries of Education in South Tyrol, Tyrol, and Thuringia.

References

ABEL A., ANSTEIN S. (2010). Korpus Südtirol - Varietätenlinguistische Untersuchungen. In : ABEL A., ZANIN R. (eds.). *Korpusinstrumente in Lehre und Forschung*, Bozen : Bozen University Press.

ABEL A., ANSTEIN S., PETRAKIS S. (2009). Die Initiative Korpus Südtirol. In : *Linguistik online* 38, 2.

ABEL A., VETTORI C., WISNIEWSKI K. (in preparation). *Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Bozen : Eurac.

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

AMMON U. (2005). Pluricentric and divided languages. In : AMMON U. ET AL. (eds.). *Sociolinguistics: An International Handbook of the Science of Language and Society*, Berlin/New York : de Gruyter, pp. 1536-1543.

AMMON U. ET AL. (2004). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Berlin/New York : de Gruyter.

ANSTEIN S. (2009). Vis-À-Vis - a System to Compare Variety Corpora. In : MAHLBERG M, GONZÁLEZ-DÍAZ V., SMITH C. (eds). *Proceedings of the Fifth Corpus Linguistics Conference*, http://ucrel.lancs.ac.uk/publications/cl2009.

AUGST G., FAIGEL P. (1986). *Von der Reihung zur Gestaltung. Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13-23 Jahren.* Frankfurt : Peter Lang.

AUTONOME PROVINZ BOZEN-SÜDTIROL (ed.) (2006). *Das neue Autonomiestatut*. Bozen : Landespresseamt.

BARONI M., BERNARDINI S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3), pp. 259–274.

BECKER-MROTZEK M., BÖTTCHER I. (2006). *Schreibkompetenz entwickeln und beurteilen.* Berlin : Cornelsen.

CHRIST O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. *Proceedings of COMPLEX 1994*, pp. 23-32.

DÜRSCHEID C., WAGNER F., BROMMER S. (2010). *Wie Jugendliche schreiben. Schreibkompetenz und neue Medien.* Berlin : de Gruyter.

EGGER K. (1979). Morphologische und syntaktische Interferenzen an der deutsch-italienischen Sprachgrenze in Südtirol. In : STURE URELAND P. (ed.). *Standardsprache und Dialekte in mehrsprachigen Gebieten Europas.* Tübingen : Niemeyer, pp. 55-104.

FIX M., MELENK H. (2000). *Schreiben zu Texten – Schreiben zu Bildimpulsen. Das Ludwigsburger Aufsatzkorpus mit 2300 Schülertexten, Befragungen und Bewertungen auf CD-ROM.* Baltmannsweiler : Schneider Verlag Hohengehren.

GIACOMOZZI L. (1982). Dialektbedingte Schwierigkeiten von Schülern aus dem Südtiroler Unterland. Ergebnisse einer Fehleranalyse. In : EGGER K. (ed.). *Dialekt und Hochsprache in der Schule.* Bolzano : Athesia, pp. 75-110.

GRANGER S., HUNG J., PETCH-TYSON S. (eds.) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia : Benjamins.

GRIES S. T. (2007). Exploring variability within and between corpora: Some methodological considerations. *Corpora*, pp. 109–151.

JANDA R. D., JOSEPH B. D. (eds) (2004). *The Handbook of Historical Linguistics*. Maldon, MA : Blackwell.

LÜDELING A., KYTÖ M. (eds.) (2009). *Corpus Linguistics. An International Handbook.* Berlin : Mouton de Gruyter.

MARGEWITSCH E. (2006). *Formelhafter Sprachgebrauch in Schülertexten*. Oldenburg : Didaktisches Zentrum.

NELSON G. (2006). The Core and Periphery of World Englishes: A corpus-based exploration. *World Englishes* 25(1), pp. 115-129.

NETZEL R., PEREZ-IRATXETA C., BORK P., ANDRADE M. A. (2003). The way we write. *EMBO reports*, 4(5), pp. 446–451.

NUSSBAUMER M., SIEBER P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In : SIEBER P. (ed.). *Sprachfähigkeiten – Besser als ihr Ruf und nötiger denn je! Ergebnisse und Folgerungen aus einem Forschungsprojekt.* Aarau : Sauerländer, pp. 141-186.

REIS M. (2003). On the Form and Interpretation of German Wh-Infinitives. *Journal of Germanic Linguistics* 15.2, pp. 155-201.

RIEDMANN G. (1972). *Die Besonderheiten der deutschen Sprache in Südtirol.* Mannheim : Biblographisches Institut.

RIZZO-BAUR H. (1962). *Die Besonderheiten der deutschen Schriftsprache in Österreich und Südtirol.* Mannheim : Biblographisches Institut.

SABEL J. (2006). Impossible Infinital Interrogatives and Relatives. In : BRANDT P., FUSS E. (eds.). *Grammar, Form and Function.* Berlin : Akademie-Verlag, pp. 242-254.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.*

SCHNEIDER W. (2001). Schulleistungen im Bereich der muttersprachlichen Bildung. In : WEINERT F. (ed.). *Leistungsmessungen in Schulen.* Weinheim : Belz, pp. 143-152.

SIEBER P. (1998). *Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit.* Tübingen : Niemeyer.

SPIEKERMANN H. (2007). Standardsprache im DaF-Unterricht: Normstandard – nationale Standardvarietäten – regionale Standardvarietäten. *Linguistik online* 32, 2.