

# Vis-À-Vis - a system to compare variety corpora

Stefanie Anstein

Institute for Specialised Communication and Multilingualism

European Academy Bolzano

sanstein@eurac.edu

## Abstract

In this work-in-progress report, the overall architecture and workflow of the system Vis-À-Vis for the systematic comparison of varieties on the basis of corpora will briefly be presented. This toolkit is designed to semi-automatically extract ‘candidates’ for the varieties’ peculiarities on different levels of linguistic description in order to support and reduce linguists’ manual work combining quantitative methods with qualitative ones.

## 1. Introduction

Varieties of pluricentric languages such as English, French, or German (cf. Ammon, 1997) are each investigated and described to different extents, where the ‘main’ varieties usually have a much better NLP coverage than their lesser-used counterparts. These varieties usually have many similar characteristics, which is why it is crucial to extract the small differences that are relevant e.g. for variant lexicography (cf. Ammon et al., 2004), for standardising lesser-used varieties, or for aiding language teaching and learning.

The variety started from in this project is used in the Autonomous Province of Bolzano - South Tyrol in Northern Italy, which is a bilingual German-Italian region. Looking at the lexical level, differences usually consist in one-to-one equivalents such as *Kondominium* (*apartment building*) in South Tyrolean German vs. *Mehrfamilienhaus* in Germany, or in many-to-one equivalents such as *provisorische Ausfahrt* (*temporary exit*) vs. *Behelfsausfahrt*. More complex phenomena such as differing collocations (e.g. *weißer Stimmzettel* vs. *ungültiger Stimmzettel* for *void ballot*) or subtle semantic differences up to pragmatic particularities of a variety are more difficult to extract, e.g. the additional meaning of *Mobilität* in South Tyrol, which is a kind of *unemployment* in addition to *mobility*.

## 2. Background and motivation

To identify relevant particularities of varieties and differences between them, corpora serve as a valuable basis. Comparable corpora are being compiled in projects such as the well-known ‘International Corpus of English’ (ICE<sup>1</sup>), ‘Trésor de la Langue Française Informatisé (au Québec)’<sup>2</sup>, ‘Corpus del Español’<sup>3</sup>, or are e.g. used for work on Portuguese varieties by Bacelar do Nascimento et al. (2006). Also for German, an initiative of research centres in Basel<sup>4</sup>, Berlin<sup>5</sup>, Bolzano<sup>6</sup>, and Vienna<sup>7</sup> called ‘C4’<sup>8</sup> is developing variety corpora which are comparable with respect to contents and size.

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. Kilgarriff, 2001; Gries, 2007). Automatic filtering of statistically produced lists containing suggested ‘candidates’ for differences or peculiarities reduces manual work and supports experts in their evaluation. In such a process, trivial characteristics of a variety or knowledge already investigated and confirmed (e.g. collections of proper names or regionalisms) should be automatically removed from candidate lists. Experts can then concentrate on the evaluation and interpretation of the remaining, mostly new phenomena.

Vis-À-Vis is being developed in a doctoral thesis<sup>9</sup> in the framework of the initiatives ‘Korpus Südtirol’ and ‘C4’. A description of the first approaches towards such a comprehensive system can be found in Abel and Anstein (2008).

Related work has 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 learner language (cf. Netzel et al., 2003), to name but a few. Many of the earlier studies were conducted manually and often for very specific phenomena. In addition, there are now more statistical approaches to data extraction from parallel or even from unrelated monolingual corpora (e.g. Nazar, 2008). Several similar and also, experimentally, some completely different approaches will be looked at for helpful insights of various research directions.

### 3. System sketch

As input to Vis-À-Vis, users give the corpora to be compared, which are then annotated with standard tools (e.g. the TreeTagger; Schmid, 1994). 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. In addition, also lesser-used languages often do have specific word lists that can be given to the system to be exploited, such as place or person names available from maps or statistics offices.

In the following 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, where the details are to be decided on. First, lexical frequency statistics are applied, since the lexical level is the first and most promising linguistic area to explore. Also earlier work on finding German variants (resulting in the dictionary by Ammon et al., 2004, and in the work by Abfalterer, 2007) will be systematically checked and enhanced. Further studies will elaborate on collocations (e.g. including a system as described in Heid and Ritz, 2005) and phrases, up to more subtle semantic (e.g. using Semantic Vectors; cf. Widdows and Ferraro, 2008<sup>10</sup>) or pragmatic differences.

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. The data will be presented in a user-friendly and intuitive way to facilitate the interpretation and further processing; 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 can again be used for the annotation of approved special vocabulary or more complex phenomena in other corpora. Vis-À-Vis will be evaluated mainly with German varieties, whereas the resulting system aims to be language-independent as far as possible. In figure 1, the preliminary overall Vis-À-Vis architecture and workflow is demonstrated.

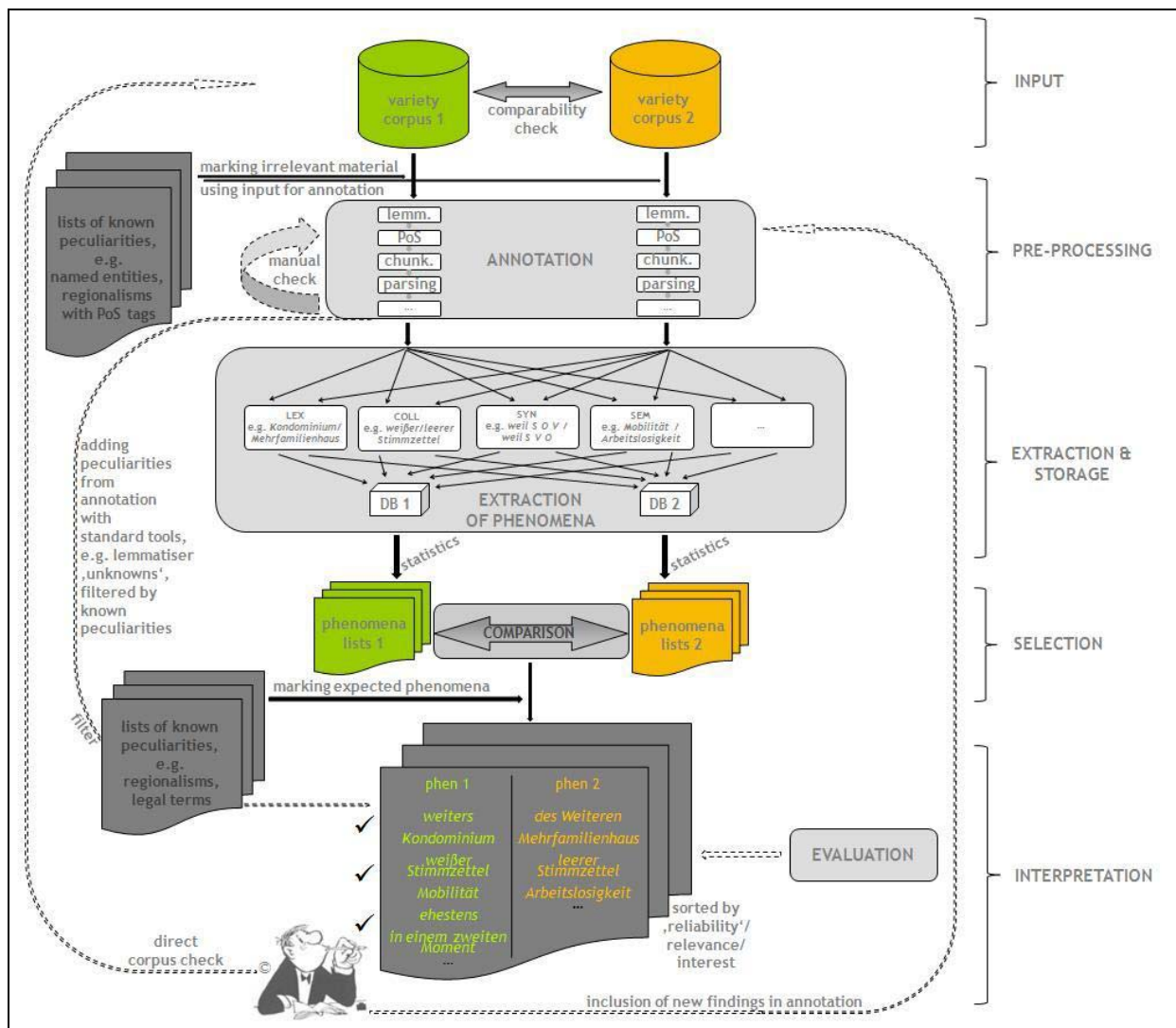


Figure 1: Vis-À-Vis overall architecture as intended

#### 4. Conclusion

This paper briefly describes ongoing work on an integrated system to compare language varieties on the basis of corpora, where manual expert work is to be supported semi-automatically. Single modules have been implemented so far; detail solutions are still to be decided on to produce a comprehensive system with adequate input and output facilities. After a comprehensive evaluation, the system will be made available to the scientific community.

## References

- Abel, A. and S. Anstein (2008). "Approaches to Computational Lexicography for German Varieties". In *Proceedings of the XIIIth Euralex International Congress*, 251–260.
- Abfalterer, H. (2007). *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*. Innsbruck: Innsbruck University Press.
- Ammon, U. (1997). *Nationale Varietäten des Deutschen*. Studienbibliographien Sprachwissenschaft, Heidelberg: Julius Groos.
- Ammon, U., H. Bickel, J. Ebner, R. Esterhammer, M. Gasser, L. Hofer, B. Kellermeier-Rehbein, H. Löffler, D. Mangott, H. Moser, R. Schläpfer, M. Schloßmacher, R. Schmidlin and G. Vallaster (2004). *Variante Wörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin/New York: Walter de Gruyter.
- Bacelar do Nascimento, M. F., J. B. Gonalves, L. Pereira, A. Estrela, A. Pereira, R. Santos and S. M. Oliveira (2006). "The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1791–1794.
- Baroni, M. and S. Bernardini (2006). "A new approach to the study of translationese: Machine-learning the difference between original and translated text". *Literary and Linguistic Computing*, 21(3), 259–274.
- Gries, S. T. (2007). "Exploring variability within and between corpora: Some methodological considerations". *Corpora*, 109–151.
- Heid, U. and J. Ritz (2005). "Extracting collocations and their contexts from corpora". In J. Pajzs et al. (eds). *Papers in computational lexicography*, 107–121.
- Janda, R. D. and B. D. Joseph (eds) (2004). *The Handbook of Historical Linguistics*. Blackwell.
- Kilgarriff, A. (2001). "Comparing corpora". *International Journal of Corpus Linguistics*, 6(1), 1–37.
- Nazar, R. (2008). "Bilingual terminology acquisition from unrelated corpora". In *Proceedings of the XIIIth Euralex International Congress*.
- Netzel, R., C. Perez-Iratxeta, P. Bork and M. A. Andrade (2003). "The way we write". *EMBO reports*, 4(5), 446–451.
- Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of International Conference on New Methods in Language Processing*.

Available at: <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger>  
(accessed: 17 September 2009)

Widdows, D. and K. Ferraro (2008). "Semantic vectors: a scalable open source package and online technology management application". In *Proceedings of the Sixth International Language Resources and Evaluation*, 1183-1190.

---

<sup>1</sup> <http://ice-corpora.net/ice>; see also bibliography list for specific variety studies

<sup>2</sup> <http://atilf.atilf.fr/tlf.htm>

<sup>3</sup> <http://www.corpusdelespanol.org>

<sup>4</sup> Schweizer Text Korpus; <http://www.schweizer-textkorpus.ch>

<sup>5</sup> DWDS-Corpus; <http://www.dwds.de>

<sup>6</sup> Korpus Südtirol; <http://www.korpus-suedtirol.it>

<sup>7</sup> Austrian Academy Corpus; <http://www.aac.ac.at>

<sup>8</sup> <http://www.korpus-c4.org>

<sup>9</sup> IMS, Stuttgart; advisor: Ulrich Heid

<sup>10</sup> <http://code.google.com/p/semanticvectors>