

VIS-À-VIS - a System for the Comparison of Linguistic Varieties on the Basis of Corpora

Stefanie Anstein

Main varieties of languages are usually far better investigated and described than their lesser-used counterparts. Even though varieties of a language in most cases do have similar linguistic characteristics, there are still differences that are to be extracted e. g. for variant lexicography (cf. Ammon et al., 2004), for standardising lesser-used varieties, or for aiding language teaching and learning. Differences on the lexical level 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’) and ‘Behelfsausfahrt’, respectively. More complex phenomena such as differing collocations or subtle semantic differences are more difficult to find, e. g. the additional meaning of ‘Mobilität’ in South Tyrol, which is a kind of ‘unemployment’ in addition to ‘mobility’.

Corpora for the varieties of a language are a valuable basis for finding relevant differences. They are being compiled in projects such as the ‘International Corpus of English’ (ICE¹) 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 Basel², Berlin³, Bolzano⁴ and Vienna⁵ called ‘C4’ is developing variety corpora comparable with respect to contents and size. Related work has been done on the comparison of language over time, of originals and translations, of native and learner language, etc. Many of these studies were conducted manually for very specific phenomena. In addition, there are more statistical approaches to data extraction from parallel or even from unrelated monolingual corpora (e. g. Nazar, 2008).

For a systematic and comprehensive comparison of corpora on different levels of linguistic description, semi-automatic tools are needed. Manual work has to be supported and reduced by the automatic filtering of statistically produced lists containing suggested ‘candidates’ for differences or peculiarities. Trivial characteristics of a variety or knowledge already investigated and confirmed (e. g. collections of proper names or regionalisms such as the above mentioned ‘Kondominium’) can be automatically removed from such candidate lists. Experts can then concentrate on the evaluation and

¹<http://www.ucl.ac.uk/english-usage/ice>

²<http://www.schweizer-texkorporus.ch>

³<http://www.dwds.de>

⁴<http://www.korpus-suedtirol.it>

⁵<http://www.aac.ac.at>

interpretation of the remaining, new phenomena, which will always have to be done manually.

The toolkit VIS-À-VIS is being developed in a doctoral thesis in the framework of the project ‘Korpus Südtirol’. It will be evaluated mainly with German varieties, but the resulting system will be language independent. The tools aim at providing support to linguists for the systematic comparison of varieties on the basis of corpora. This support consists in methods to filter huge amounts of data and to present to the expert only probably relevant material. With this approach, less manual work is necessary and quantitative methods can be combined with qualitative ones. In addition, the data to be evaluated manually are presented in a user-friendly and intuitive way to facilitate the interpretation and further processing.

As input to VIS-À-VIS, users give the corpora to be compared as well as, if available, lists with previous knowledge as described above. The corpora are then annotated with standard tools, which is where difficult cases for the tools or errors produced by them can identify the first set of candidates for special variety characteristics, since the tools are usually created for the main varieties. In the following modules, the corpora are analysed and compared with a combination of existing as well as new or adapted tools for e. g. concordancing or frequency statistics. The lexical level is the first and most promising linguistic area to explore; further studies will elaborate on collocations and phrases up to more subtle semantic or pragmatic differences. The knowledge about the variety is taken into account in all the modules wherever possible. As a result, VIS-À-VIS produces filtered lists of probably relevant differences between the varieties for manual evaluation. It is also possible for the user to search directly in the relevant corpora for sentence contexts of ambiguous or other difficult cases. In a further step, the findings can again be used for the annotation of approved special vocabulary or more complex phenomena in other corpora of that variety to be compared.

A description of the first approaches to this toolkit can be found in Abel and Anstein (2008). In this poster, the overall VIS-À-VIS architecture and workflow is to be demonstrated. Since ongoing work is being described, it includes a discussion on possible alternative detail solutions and future work.

References

- Abel, A. and Anstein, S. (2008). Approaches to Computational Lexicography for German Varieties. In *Proceedings of the XIIIth Euralex International Congress*, pages 251–260, Barcelona.
- Ammon, U., Bickel, H., Ebner, J., Esterhammer, R., Gasser, M., Hofer, L., Kellermeier-Rehbein, B., Löffler, H., Mangott, D., Moser, H., Schläpfer, R., Schloßmacher, M., Schmidlin, R., and Vallaster, G. (2004). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Walter de Gruyter, Berlin/New York.
- Bacelar do Nascimento, M. F., Gonalves, J. B., Pereira, L., Estrela, A., Pereira, A., Santos, R., and Oliveira, S. M. (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 (LREC)*, pages 1791–1794, Genoa, Italy.
- Nazar, R. (2008). Bilingual terminology acquisition from unrelated corpora. In *Actas del XIII Congreso Internacional Euralex*. European Association for Lexicography, Universidad Pompeu Fabra.