Comparing Variety Corpora with *Vis-À-Vis* – a Prototype System Presentation

Stefanie Anstein
Institute for Specialised Communication and Multilingualism
European Academy of Bozen/Bolzano (EURAC)
Viale Druso 1
I – 39100 Bolzano
stefanie.anstein@eurac.edu

Abstract

In this paper, the prototype system *Vis-À-Vis* to support linguists in their comparison of regional language varieties is presented. Written corpora are used as an empirical basis to extract differences semi-automatically. For the analysis, existing and adapted as well as new tools with both pattern-based and statistical approaches are applied. The processing of the corpus input consists in the annotation of the data, the extraction of phenomena from different levels of linguistic description, and their quantitative comparison for the identification of significantly different phenomena in the two input corpora. *Vis-À-Vis* produces sorted 'candidate' lists for peculiarities of varieties by filtering according to statistical association measures as well as using corpus-external knowledge to reduce the output to presumably significant phenomena. Traditional regional variety linguists benefit from these results using them as a compact empirical basis – extracted from large amounts of authentic data – for their detailed qualitative analyses. Via a user-friendly application of a comprehensive computational system, they are supported in efficiently extracting differences between varieties e. g. for documentation, lexicography, or didactics of pluri-centric languages.

1 Background and related work

Pluri-centric languages are languages with more than one national center and with specific national varieties (Clyne, 1992). The latter usually differ to a certain extent on different levels of linguistic description, mostly on the lexical level – an example for variants in German being *Marille* (used in Austria and South Tyrol) vs. *Aprikose* (used in Germany and Switzerland) for 'apricot'.

The question to be answered in the framework research project is to what extent the comparison of varieties for supporting variety linguists' manual analyses can be automated with natural language processing (NLP) methods. The analysis results obtained with such computational systems will contribute to variety documentation, lexicography, and language didactics.

Vis-À-Vis has been developed for the case of the pluri-centric language German (Ammon, 1995) for the time being; its development originated in the initiatives *Korpus Südtirol*[1] and *C4*[2]. The former is preparing a written text corpus of South Tyrolean German[3] (Anstein et al., 2011), which can also be queried together with other German variety corpora with the help of the distributed query engine implemented in the *C4* project (Dittmann et al., 2012). In addition to interactively run single queries in the *C4* corpora, variety linguists can use *Vis-À-Vis* to exploratively and empirically analyse and compare corpora on the desired levels of linguistic description. This is especially relevant since the amount of electronically available data constantly increases and can no longer be handled purely manually. The benefit of supportive tools from the NLP community for em-

[1] http://www.korpus-suedtirol.it

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

[3] South Tyrolean German is the German variety used as an official language in the Autonomous Province of Bolzano / South Tyrol in Northern Italy (Egger and Lanthaler, 2001).

pirical analyses in traditional linguistics is clearly evident in this scenario, which is where the usefulness of this approach can be seen.
Other work that is related to this topic has been done in general comparative corpus linguistics as e. g. described in McEnery et al. (2006) or in Schmied (2009). Comparative regional variety linguistics has first been handled mostly manually with single introspective studies or later as well with the help of the Internet, e. g. in the development of the *Variant Dictionary of German* (Ammon et al., 2004). By now, more and more projects use variety corpora and automated comparison methods, e. g. the *ICE*[4] initiative or Bacelar do Nascimento et al. (2006) studying the varieties of Portuguese.

2 The system *Vis-À-Vis*

In this section, the toolkit's implementational and functional details as well as its accessibility are described.

2.1 Design and implementation

Vis-À-Vis is written in the programming language *Perl*[5] with a modular approach.

Input and output The main script takes as input (i) written text corpora of two varieties and, if available, (ii) lists with known peculiarities (e. g. named entities or regionalisms) of the variety to be investigated with respect to the so-called reference variety. The output is composed by (i) general quantitative information on the corpora and their comparability as well as (ii) lists of phenomena occurring in the two corpora, sorted by frequency and statistical values including filtering information for identifying new regionalism candidates.

Architecture As a first step, the two input corpora are checked with regard to their comparability. Then the corpora are annotated including corpus-external linguistic knowledge. In the extraction module, phenomena from different levels of linguistic description are identified. These are further compared by frequency and by statistical association measures and are presented to the user together with filter information for their interpretation. The overall *Vis-À-Vis* design can be seen in figure 1; details of the modules are given in section 2.2.

Approaches Both top-down / corpus-based and bottom-up / corpus-driven methods are applied; the former for the revision of possibly existing, manually compiled variant lists and the latter for their enhancement. Morpho-syntactic patterns according to part-of-speech (PoS) tags are used as well as explorative statistical approaches on the basis of significance measures.

2.2 Functionalities

In the following, the system's functional features are elaborated on.

Comparability check As a measurement for the comparability of the two corpora to judge the reliability of the comparison results (see also Gries, 2007), their 'complexity', as also investigated e. g. in learner corpus studies, is taken. On the one hand, the type-token ratio is calculated, which is an indicator for vocabulary richness and lexical variability. On the other hand, the proportion of lexical to grammatical words is given, measured as lexical density (Stubbs, 1986).

Annotation After tokenisation, the corpora are PoS-tagged and lemmatised with the *TreeTagger*[6]. The corpus-external lexical lists are used to lemmatise words that are not known to the tagger. In a bootstrapping process, new findings can be integrated via the annotations into a new *Vis-À-Vis* run by providing new lexicon entries as additional input.

Analysis levels On the lexical level, all word forms or lemmas of the two corpora are counted with the *Corpus Query Processor* (*CQP*)[7]. On the bi-gram level, the extraction of co-occurrences is done by searching for PoS patterns (e. g. adjective + noun or adverb + adjective) via *CQP* corpus queries. On an exemplary higher level of linguistic description, frequencies of main and subordinate clauses for both corpora are provided and the

[4] International Corpus of English; `http://ice-corpora.net/ice/index.htm`; Nelson (2006)

[5] `http://www.perl.org`

[6] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`; Schmid (1994)

[7] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`; Christ (1994)

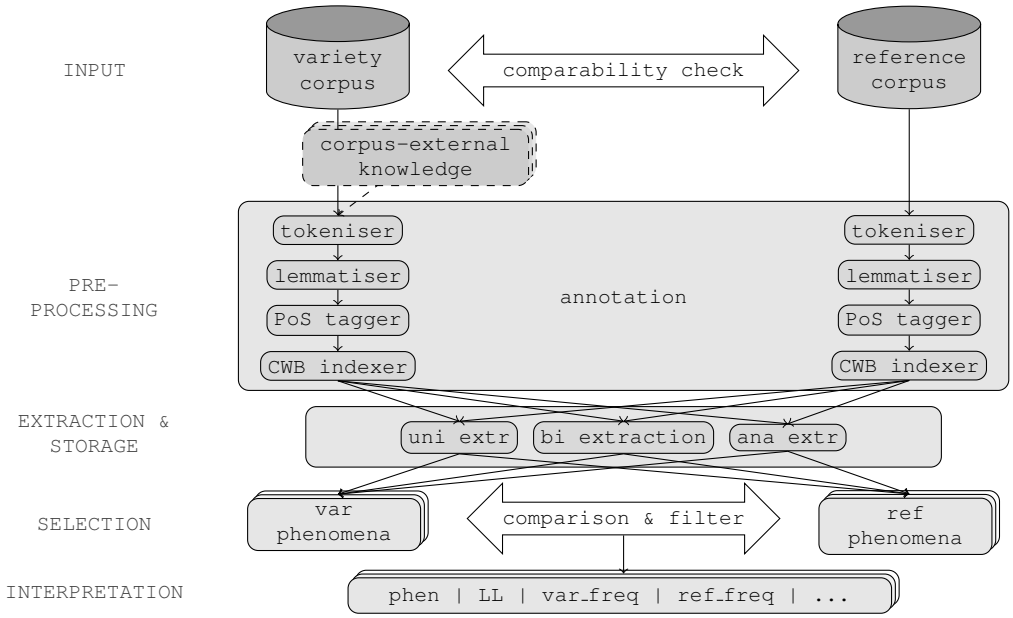

Figure 1: Overall architecture of *Vis-À-Vis*

word order in subordinate clauses is investigated. The extraction is done by *CQP* queries for specific PoS patterns, e. g. a subordinating conjunction with verb-second word order,which is an anacoluthon (sentence 'break') in written language.

Comparison and statistics The difference of phenomenon occurrences in the two corpora is determined (i) with absolute as well as relative frequencies with respect to corpus sizes and (ii) with statistical association measures. Two measures indicate how significantly different the frequency of one phenomenon is in the variety corpus with respect to the reference corpus. The `log-likelihood (LL)` measure (Dunning, 1993) was chosen as an association measure recommended e. g. by Rayson and Garside (2000) for co-occurrences and also by Evert (2004) both for words and collocations. As a second measure, `LL*log(frequency)` is given, since Kilgarriff and Tugwell (2002) state that `LL` values over-emphasise the significance of low-frequency items and thus suggest to adjust these values for measuring e. g. lexicographic relevance.

Filtering The output lists are marked according to several external knowledge lists, partly system-internal and, if available, also provided by the user. They consist of known regionalism lists (e. g. 'Südtirolisms'[8] taken from Abfalterer, 2007), place and person name lists, and lists of 'reality' descriptions (Heid, 2011) such as currency names. By filtering out known information, new regionalism candidates can be identified by sorting the output lists accordingly. In addition, the statistical measures and a filter according to expected frequency values serve as a guideline to the probability of the candidates to be relevant.

2.3 Access and extensibility

The system will be released with a free software license for the download as a stand-alone application, and it can also be used online from the *Korpus Südtirol* website. The download comprises two possibilities of usage – Unix command line use and a graphical user interface (GUI) for both Unix and Windows environments. A comprehensive system documentation with all details for its usage is provided for all scenarios.

Command line use The script `visavis.perl` supports several parameters as described in the following. With the option `-f`, the user decides

[8] Südtirolisms ('South-Tyrol-isms') are the specific variants of linguistic units used in South Tyrolean German. The terms 'primary' (exclusively used in South Tyrol) and 'secondary' (shared with other varieties) Südtirolisms have been coined by Abfalterer (2007).

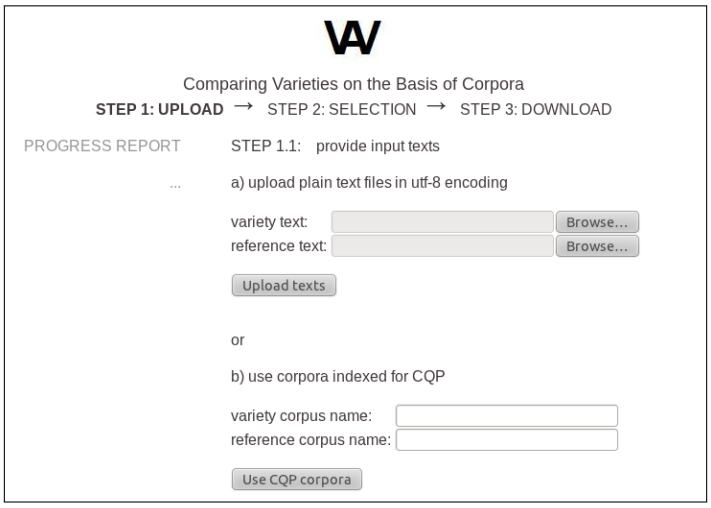

Figure 2: *Vis-À-Vis* GUI start page – corpus upload

if either word forms or lemmas are to be considered in the analysis process. The option `-l` chooses the level of extraction and comparison (lexical, co-occurrence, or anacolutha). The option `-e` takes corpus-external knowledge for the variety of all kinds as input, lexicon entries (one-word units) with PoS and lemma information and external knowledge lists with certain prefixes to each of their filenames. Finally, the option `-i` is used to specify if the two input corpora are in text format or if they are available corpora indexed for *CQP*. As results, the user gets general data, e. g. regarding the comparability of the two corpora, as well as the locations of the comparison output files to view or further process printed in the terminal window.

Graphical user interface For easier accessibility of *Vis-À-Vis*, users can upload their data over a GUI and are guided through the options for the comparison process up to the download of their analysis results. In figure 2, the start page of the *Vis-À-Vis* GUI is shown by a screenshot to give an idea of its layout. After choosing the kind of corpus input and providing the data locations, lists with corpus-external knowledge can be uploaded, if available. In the second step, the desired analysis and comparison level is chosen, and after the *Vis-À-Vis* run, the result data can be viewed and downloaded for further processing. Through the GUI, also a direct link to *Korpus Südtirol* for the verification of South Tyrolean phenomena and for context search is provided.

Extensions In the stand-alone version, several possibilities to adapt and extend the system are given, for example: integration of additional annotation, extraction, and comparison tools, usage of other comparability measures, extension of the analysis levels, enhancement of the filtering, application of additional statistical measures for comparison, or adaptation to other languages. Also the integration of the tool into a larger corpus processing architecture is a possible and promising development to be followed further.

3 Evaluation and conclusion

The system evaluation using Abfalterer's Südtirolisms as a gold standard showed promising results for *Vis-À-Vis*' approach. First concrete outcomes obtained by using *Vis-À-Vis* for lexicographic tasks are new as well as refined dictionary entries for South Tyrolean German, e. g. for the lemmas *ehestens* (as soon as possible), *Konsortium* (consortium), *ober* (above), or *weiters* (furthermore); for details see Abel and Anstein (2011). Detailed precision, recall, and f-score values on the basis of different parameters can be found in Anstein (to appear).

Given such findings, it seems worth following *Vis-À-Vis*' approach and develop it further to provide an even more useful NLP tool for traditional linguistics, also to serve in other fields than regional variety linguistics – wherever corpora are to be compared in order to find significantly different phenomena.

References

Andrea Abel and Stefanie Anstein. 2011. Korpus Südtirol - Varietätenlinguistische Untersuchungen. In Andrea Abel and Renata Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, Bolzano. University Press.

Heidemaria Abfalterer. 2007. *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht: lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*, volume 72 of *Germanistische Reihe*. Innsbruck University Press.

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

Ulrich Ammon. 1995. *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. De Gruyter, Berlin / New York.

Stefanie Anstein, Margit Oberhammer, and Stefanos Petrakis. 2011. Korpus Südtirol - Aufbau und Abfrage. In Andrea Abel and Renata Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, Bolzano. University Press.

Stefanie Anstein. to appear. Computational Approaches to the Comparison of Regional Variety Corpora – Prototyping a Semi-automatic System for German. PhD thesis.

Maria Fernanda Bacelar do Nascimento, José Bettencourt Gonçalves, Luísa Pereira, Antónia Estrela, Alfonso Pereira, Rui Santos, and Sancho M. Oliveira. 2006. The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon. In *Proceedings of the Fifth International Language Resources and Evaluation Conference (LREC 2006)*, pages 1791–1794, Genoa, Italy.

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX)*, pages 23–32, Budapest.

Michael G. Clyne, editor. 1992. *Pluricentric languages: Differing norms in different nations*. De Gruyter, Berlin / New York.

Henrik Dittmann, Matej Ďurčo, Alexander Geyken, Tobias Roth, and Kai Zimmer. 2012. Korpus C4 – a distributed corpus of German varieties. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, number 14 in Hamburg Studies in Multilingualism (HSM). John Benjamins, Amsterdam.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1).

Kurt Egger and Franz Lanthaler, editors. 2001. *Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt*. Folio, Vienna / Bolzano.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Ulrich Heid. 2011. Korpusbasierte Beschreibung der Variation bei Kollokationen: Deutschland – Österreich – Schweiz – Südtirol. In Stefan Engelberg, Anke Holler, and Kristel Proost, editors, *Sprachliches Wissen zwischen Lexikon und Grammatik*, Jahrbuch 2010. De Gruyter, Institut für Deutsche Sprache, Mannheim.

Adam Kilgarriff and David Tugwell. 2002. Sketching words. *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137.

Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies - An Advanced resource book*. Routledge Applied Linguistics. Routledge.

Gerald Nelson. 2006. The core and periphery of world englishes: a corpus-based exploration. *World Englishes*, 25(1):115–129.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester.

Josef Schmied. 2009. Contrastive corpus studies. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, number 29 in Handbooks of Linguistics and Communication Science, chapter 54, pages 1140–1159. De Gruyter, Berlin.

Michael Stubbs. 1986. Lexical density: A computational technique and some findings. In Malcolm Coultard, editor, *Talking about Text. Studies Presented to David Brazil on His Retirement*, pages 27–42. English Language Research, Birmingham.