News from the STyrLogism Project: Language varieties meet One-Click Dictionary

Egon Stemle, Eurac Research, Bozen

Abstract

The goal of the *STyrLogism Project* is to semi-automatically extract neologism candidates for the German standard variety used in South Tyrol (STyrGerman), an autonomous province in Northern Italy characterised by a multilingual environment with three official languages (Italian, German, and Ladin), an institutional bilingualism or trilingualism (depending on the percentage of the Ladin population), and diverse individual language repertoires (Ammon 1995).

Immediate use-cases for these neologisms include, for example, consideration for future editions of the "Variantenwörterbuch des Deutschen" (Dictionary of variants of the German language, abbr. VWB) (Ammon, Bickel, and Lenz 2016) and other dictionaries; furthermore, the language centre of the South Tyrolean cultural institute[1] observes trends in the local variety of German, and is interested in supplementing their activity with data from neologism candidates; or, more generally, the project can be used as an empirical basis for the long-term observation and evaluation of trends of the local standard variety of the German language, which makes it interesting for language policy and language planning measures.

So far, the VWB has taken variants of the German language only into account if the corresponding words often appear in standard texts, i.e. usually newspaper texts. But standard texts alone do not unequivocally cover the entire relevant language usage. For example, the meaning of “Bar” in STyrGerman (“a place to have coffee”, i.e. a “coffee shop”, and not “a place to have drinks, especially in the evening”, i.e. a “night bar”, like in other varieties of German) is usually not conveyed in newspaper texts. There, “Bar” is often mentioned, for example, together with break-ins, but is hardly described in a way to infer its different usage (Abel 2018). Many relevant linguistic phenomena can be monitored not only with standard text corpora but additionally – and some phenomena even better – with web corpora and corpora of computer-mediated communication.

We use a list of manually vetted web addresses (URLs) from news, magazines and blog sites in South Tyrol to crawl their data with the Internet Archive’s open-source web crawler Heritrix[2] and save the content in Web ARChive (WARC) format[3]. We then construct our corpus with (parts of) the COW14 architecture (Schäfer 2015), which performs basic cleanups and boilerplate removal, simple connected text detection as well as a two-step dupulicate detection: the first step removes perfect duplicates, i.e. documents that are identical up to the last character; the second step removes near-duplicates by calculating token-n-grams for each page and the corresponding fingerprint (w-shingle). This fingerprint has the property that pages with similar

[1] https://kulturinstitut.org/sprachstelle/
[2] https://github.com/internetarchive/heritrix3/wiki
[3] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/

content result in similar fingerprints, so the data can be de-duplicated by selecting a range of similarity values between the fingerprints.

The resulting data is converted into a list of word forms and a corpus for the NoSketchEngine (NoSkE) (Rychlý 2007). We then do case-insensitive comparisons of the list of word forms with a) those from our reference corpora, b) the additional word lists, which is ultimately a simple Named Entity Recognition, and c) with the combination of all formerly crawled data sets. Our reference corpora are DECOW14 (Schäfer & Bildhauer 2012) with around 60 million word forms, and the South Tyrolean Web Corpus (Schulz, Lyding, and Nicolas 2013) with around 2.4 million word forms; the additional word lists consist of proper names and terminological terms from the region as well as specific terms of STyrGerman (a total of around 53,000 word forms). The cleaned data is then tokenised – but not lemmatised – and converted into a word list.

The list of *neologism candidates* then consists of the word forms in a crawl that occur less than a predefined number of times in all previous data. Next, the candidates are checked manually in a specifically designed, optimised interface. This interface shows a certain number of neologism candidates along with their first (and possibly only) result in a KWIC view. The user can then click on a candidate to get the entire result page of that candidate's search query in the NoSkE, where additional meta information is available. The user can also click a checkbox or enter a comment into a text field (which automatically triggers the checkbox) to mark this candidate for later curation. Finally, the user can proceed to the next page and automatically discard all unmarked candidates on the current page from further processing.

Here, we will report on our previous work (Abel and Stemle 2018) and will relate it to our current work that is conducted as part of our institution's observer status in the European Lexicographic Infrastructure (ELEXIS) project (Simon Krek et al. 2018). ELEXIS features the *One-Click Dictionary* tool chain to automatically generate, for example, headword lists, word (and other lexical unit) senses, definitions, and corpus based examples. The tool chain consists of the corpus query system Sketch Engine (Kilgarriff et al. 2014) and the dictionary writing system Lexonomy (Měchura 2017); together they are supposed to support lexicographers along the entire pipeline of producing a dictionary, from corpus to screen, where dictionaries are pre-generated automatically from a corpus (using the Sketch Engine) and then post-edited (using Lexonomy).

References

Abel, Andrea. 2018. "Von Bars, Oberschulen Und Weißen Stimmzetteln: Zum Wortschatz Des Standarddeutschen in Südtirol." In *Deutsch Als Minderheitensprache in Italien: Theorie Und Empirie Kontaktinduzierten Sprachwandels*, edited by Stefan Rabanus, 283–323. Germanistische Linguistik, 239-240 / 2018. Hildesheim, Germany: Georg Olms Verlag.

Abel, Andrea, and Egon W. Stemle. 2018. "On the Detection of Neologism Candidates as Basis for Language Observation and Lexicographic Endeavours: The STyrLogism Project." In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*,

edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem, and Simon Krek, 535–44. Ljubljana, SI: Ljubljana University Press, Faculty of Arts. https://doi.org/10.4312/9789610600961.

Ammon, Ulrich. 1995. *Die Deutsche Sprache in Deutschland, österreich Und Der Schweiz: Das Problem Der Nationalen Varietäten*. Berlin/New York: De Gruyter.

Ammon, Ulrich, Hans Bickel, and Alexandra Nicole Lenz, eds. 2016. *Variantenwörterbuch Des Deutschen: Die Standardsprache in österreich, Der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien Und Südtirol Sowie Rumänien, Namibia Und Mennonitensiedlungen*. 2nd ed. Berlin/Boston: De Gruyter Mouton.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years on." *Lexicography* 1 (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Měchura, Michal. 2017. "Introducing Lexonomy: An Open-Source Dictionary Writing and Publishing System." In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference*. Leiden, Netherlands.

Rychlý, Pavel. 2007. "Manatee/Bonito – A Modular Corpus Manager." In *First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, 65–70. Brno, Czech Republic: Masaryk University.

Schäfer, Roland. 2015. "Processing and Querying Large Web Corpora with the COW14 Architecture." In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, edited by Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, 28–34. Lancaster, UK. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3826/file/Schaefer_Processing_and_querying_large_web_corpora_2015.pdf.

Schulz, Sarah, Verena Lyding, and Lionel Nicolas. 2013. "STirWaC: Compiling a Diverse Corpus Based on Texts from the Web for South Tyrolean German." In *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, edited by Stefan Evert, Egon Stemle, and Paul Rayson, 35–45. Lancaster, UK. http://hdl.handle.net/1854/LU-4167988.

Simon Krek, Iztok Kosem, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, and Tanja Wissik. 2018. "European Lexicographic Infrastructure (ELEXIS)." In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem, and Simon Krek, 881–91. Ljubljana, SI: Ljubljana University Press, Faculty of Arts.