Language Varieties Meet One-Click Dictionary

Egon W. Stemle, Andrea Abel, Verena Lyding

Institute for Applied Linguistics, Eurac Research, Bolzano - Bozen, IT

E-mail: {egon.stemle,andrea.abel,verena.lyding}@eurac.edu

Abstract

The goal of the STyrLogism Project is to semi-automatically extract neologism candidates (new lexemes) for the German standard variety used in South Tyrol, and generally to create the basis for long-term monitoring of its development. We use automatic lexico-semantic analytics for the lexicographic processing, but instead of continuing to develop our independent neologism detection application, we have recently become part of a thriving community of users and developers within the EU infrastructure project ELEXIS, which aims to harmonize efforts that relate to producing and making dictionary resources available, and to develop tools with consistent standards and increased interoperability. Consequently, we moved the development of our neologism application into Lexonomy, one of ELEXIS' promoted open-source projects. In the following, we report on the current state of this ongoing development by describing how we integrate our work with the Sketch Engine and Lexonomy tools, pointing out the challenges involved, and discussing how our work on language varieties can be evaluated.

Keywords: language variety; One-Click Dictionary; web corpus; dictionary of variants; ELEXIS

1. Introduction

The goal of the STyrLogism Project is to semi-automatically extract neologism candidates (new lexemes) for the German standard variety used in South Tyrol (STyrGerman), an autonomous province in Northern Italy where German is an official language. Direct applications for these neologisms are, for example, the consideration for future editions of the *Variantenwörterbuch des Deutschen* (Dictionary of variants of the German language, abbr. VWB) (Ammon et al., 2016) and other dictionaries. In the medium term, the project should additionally serve as an empirical basis for the long-term observation and evaluation of trends in STyrGerman, which also makes it interesting for language policy and language planning measures.

In total, there are up to three official languages (Italian and German - plus Ladin, in the Ladin valleys) and an institutional bi- or trilingualism in the region, which means that the two (or three) languages have the same standing and there is an effective multilingual obligation of the civil servants and the right to address the administration in one of the languages; through personal linguistic socialization and individual biographical constellations and experiences, people in the region usually acquire diverse individual language repertoires, that is (multilingual) dynamic communicative competences. Moreover, from a pluricentric perspective, South Tyrol is a national semi-centre, inhabits a peripheral location in the German-speaking area, and offers an interesting language contact situation, especially with regard to the German and Italian languages (Abel, 2018). All this makes South Tyrol in general and STyrGerman in

particular an interesting object of investigation for linguistic studies.

The completely revised second edition of VWB appeared in 2016, 12 years after the first edition, but STyrGerman could not be analysed to the same extent as the varieties of the full centres (Germany, Austria, Switzerland) and, in addition, methodological decisions led to some developments and phenomena being less represented: Firstly, only corpus data with journalistic prose served as a source for the new edition of the VWB, while for the first edition various text genres had been used, which were not based on digital corpora yet but on text excerpts on paper. However, standard texts from newspaper corpora alone do not unequivocally cover the entire relevant language usage. For example, "*Bar*" has a particular meaning in STyrGerman in the sense that it is used to refer to a place to have coffee, that is, as a synonym for "coffee shop", whereas in the other German varieties it only has the meaning "night bar", and it is difficult to extract sentences conveying this STyrGerman meaning of "Bar" from newspaper texts. In them, "Bar" is often mentioned, for example, together with break-ins, but is hardly described in a way to infer its different usage (e.g. mentioning what people usually do there, drinking coffee, eating a croissant, reading the newspaper). A case in point is the following excerpt from original data: "Zu der Bluttat war es vor dem Eingang der 'Bar Pleres' in Matsch gekommen" (translation: "The bloody deed took place in front of the entrance of the 'Bar Pleres' in Matsch")[1] (Abel, 2018). Furthermore, 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, because language changes on social media and the internet can be in public online usage for a while before getting included into mainstream newspapers and other text genres (Androutsopoulos, 2011). However, social media and web corpora were not included in the data for the VWB.

Secondly, in the course of the VWB project, it was not possible (for financial reasons) to check systematically whether new STyrGerman lexemes should be included or obsolete ones should be eliminated. This is a matter of linguistic change that is closely related to the research on neologisms, which in our case also includes variants that are commonly used in STyrGerman but are not yet lexicalised (Abel & Stemle, 2018). We are aware that these are not neologisms in the narrower sense, but we do not need to make this distinction with regard to data processing. The research on neologisms is typically divided into two categories: one category for words used in a new meaning, and another for new lexemes with an unseen graphical representation (Kinne, 1998). In the past, we have concentrated on the detection of neologism candidates of the latter category. As an example, we can mention "Vollkornpizzetta" ("very small, round-shaped pizza made of whole-grain"). The particular part of this compound word is "Pizzetta" that derives from "Pizza" being "-etta", the diminutive suffix in Italian. But the whole word is not a loan word from Italian; the compound modifier "Vollkorn" is the German word for "whole-grain" and not the Italian word "integrale". However,

[1] Dolomitenkorpus, 2001: http://www.korpus-suedtirol.it

it would not be the same to talk about a "kleine Vollkornpizza" ("small pizza, minipizza"), because "pizzetta" in Italian refers to a particular type of pizza, usually a very small, round-shaped pizza (with a diameter of around 5 cm), which you offer, for example, at a buffet as finger food.

Lastly, the focus for including variants was on the occurrence of different word forms and on differences in word meanings, but there exist collocations which are not specific for a variety because of their individual words, but because the words are frequently combined and thus represent a collocation. For example, the meaning of "jemanden in die Mobilität entlassen/überstellen" (literal translation "*to release/transfer someone into mobility"; the actual meaning "to let someone go after a company struggled for some time" is a transfer from the Italian "mobilità") is only specific to STyrGerman (Abel, 2018).

Overall, as reported in earlier work (Abel & Stemle, 2018), the STyrLogism Project changes some of the collection parameters and attempts to remedy some of the aforementioned shortcomings. First and foremost, we use web data as a valuable complement to standard texts (Barton & Lee, 2013), so that we can now observe short-term and fast-moving developments in online media. Overall, we aim to provide semi-automatic support for the detection of new lexemes and lexeme combinations that are more frequent in STyrGerman than in other variants—or even exclusive to STyrGerman—and, finally, we also want to employ methods to detect meaning shift, which previously has been done manually as part of exploratory analyses within the project.

2. Related Work

The approaches for neologism detection can be divided into two groups. One, usually applied to a single set of new data, uses language resources such as word lists or linguistic patterns. The word lists are compiled from existing lexicographic resources such as dictionaries or corpora, combined with filters to eliminate non-words, typographical errors, named entities, and so on, and the linguistic patterns are, for example, markers of lexical novelty like punctuation marks that can signal new words, as shown in O'Donovan and O'Neill (2008) and Paryzek (2008). The other group, usually applied to multiple datasets, uses statistical measures or machine learning to calculate and evaluate the increase in usage or the change in meaning over time or in different registers. Examples can be found in Stenetorp (2010) and Kilgarriff et al. (2015). Finally, these two approaches can also be combined.

Wortwarte[2] (Lemnitzer, 2000-2019) is the most relevant previous project in relation to our own, as it is an ongoing project with an online portal that has been regularly collecting and documenting new German words. The system is based on German online-newspaper texts: a web crawler regularly collects data from pre-defined sites, such as

[2] http://www.wortwarte.de/

newspapers and magazines. After the HTML content has been cleaned up, the plain text is used to build a new time slice of a corpus. The selection of neologism candidates is based on short-term evaluations in which the new corpus is compared with the continuously growing German reference corpus (Das Deutsche Referenzkorpus – DeReKo. See Kupietz and Lüngen (2014) for an overview) with around 42 billion word tokens (status: Q1.2018). In order to avoid "random" errors (e.g. typing errors) and to filter out spelling mistakes, the selection of neologisms is conducted manually after the comparison with DeReKo. The results of these analyses are published online at irregular intervals, but typically about once a week. The results usually include a few words with their exemplary use in a sentence and the reference as to where they came from.

O'Donovan and O'Neill (2008) use a similar idea, but due to the lack of free access to a continuously growing reference corpus for English they use and update their own Chambers Harrap International Corpus (CHIC) web corpus. It consists of more than 500 million words of International English and stands in the tradition of the Bank of English rather than a static, balanced resource like the British National Corpus (BNC). They also use other resources, like lemmatization and morpho-syntactic information, such as a headword list augmented with inflected forms. Kerremans, Stegmayr, and Schmid (2011) also crawl their own reference corpus and additionally use an explicit component to monitor the changed over time for selected terms: they use the commercial search engine Google and regularly crawl the content of search results returned for each "to-be-monitored" neologism.

3. STyrLogism: Evolution

3.1 Initial implementation

The first incarnation of the STyrLogism Project system (Abel & Stemle, 2018) consisted of a list of manually selected URLs from news, magazines and blog websites of South Tyrol, and regular data crawls from the Heritrix[3] Internet Archive crawler. The whole content from the crawled web pages was saved in the Web ARChive (WARC) archive format. Then, we used Schäfer and Bildhauer's (2012) texrex toolkit. This comes already set up to process WARC files, and directly works with the Heritrix output. It removes HTML and scripts, and uses a simplistic heuristic to split paragraphs in the resulting text. So-called boilerplate, that is, navigational elements and menus, date strings, copyright notices, among others, are then identified and quantified as an annotation on a paragraph level. Finally, a two-step duplicate detection is employed: the first removes perfect duplicates, that is, documents that are identical up to the last character; the second step removes near-duplicates. The resulting data was converted into a list of word forms and a corpus for the NoSketchEngine (NoSkE) (Rychlý, 2007). We then made case-insensitive comparisons of the list of word forms with: a) the one from our reference corpora, b) the additional

[3] https://archive.org/projects/

word lists, which was in practice a simple Named Entity Recognition, and c) with the combination of all formerly crawled data sets. Our reference corpora were DECOW14 (Schäfer & Bildhauer, 2012) with around 60 million word forms, and the South Tyrolean Web Corpus (Schulz et al., 2013) with around 2.4 million word forms; the additional word lists consisted of named entities, terminological terms from the region, and specific terms of the German standard variety used in South Tyrol (altogether around 53,000 word forms). The cleaned data of the latest crawl was then tokenized—but not lemmatized—and converted into a word list. This list of candidate words consisted of those in the latest crawl that appeared less than a predefined number of times in all of the other data. Finally, the candidates were manually checked in a specifically crafted streamlined interface. This interface shows a predefined number of neologism candidates on one page along with the first (and possibly only) results as a KWIC result. The user can then click the candidate to get the whole result page of this candidate's search query in the NoSkE, where all additional meta information for each search result is available. The user can also click a check-box or enter a comment into a text field (which automatically triggers the check-box) to make a note of this candidate for later curation. Finally, the user can go to the next page, which automatically discards all unmarked candidates from further processing. In a second 'curation' step, a user can see all the previously marked candidates with single KWIC results of all occurrences of the candidate in different crawler runs. This stage gives an overview of the currently tracked neologism candidates with quick access to individual occurrences over time.

3.2 Updated Method

Here, we will report on our current work that is conducted as part of our institution's observer status in the European Lexicographic Infrastructure (ELEXIS) project (Krek et al., 2018). ELEXIS features the One-Click Dictionary toolchain to automatically generate, for example, headword lists, word (and other lexical units) senses, definitions, and corpus-based examples. The toolchain consists of the corpus query system Sketch Engine[4] (Kilgarriff et al., 2014) and the dictionary writing system Lexonomy[5] (Měchura, 2017); together they are supposed to support lexicographers along the entire pipeline of producing a dictionary (see Granger & Paquot (2012) for an overview of electronic means in the planning, writing, and dissemination of dictionaries), from corpus to screen, where dictionaries are pre-generated automatically from a corpus (using Sketch Engine) and then post-edited (using Lexonomy).

ELEXIS, among other things, aims to harmonize efforts on a larger European scale that relate to producing and making dictionary resources available, and to develop tools to update existing or new resources with consistent standards and increased interoperability. We hope that through cooperation within ELEXIS more opportunities

[4] https://www.sketchengine.eu

[5] https://www.lexonomy.eu

and desirable developments arise: With access to current methods and tools, and a collective awareness of challenges and information about upcoming solutions for the next generation of online dictionaries, we can integrate our local digital resources into modern workflows and also provide feedback that influences the design of use-cases for tools and workflows.

The One-Click Dictionary is a convenient automation for exchanging lexicographic data between a Sketch Engine corpus and a Lexonomy dictionary, and will eventually cover, for example, the extraction of example sentences, the detection of definitions, descriptions and collocations, and the clustering of word senses. The computations and analyses are carried out by the Sketch Engine, and the results are transmitted to Lexonomy as dictionary entries. The communication is channelled through an Application Programming Interface (API), that is a set of defined functions and procedures that lets computers talk to each other. In Lexonomy, the data can then be edited and eventually published as an online dictionary, ideally under an open-source license, for example, CC0, CC-BY, CC-BY-SA[6] or ODbL[7]. There will also exist some dedicated features for post-editing an automatically generated dictionary: for example, features for quickly splitting and lumping senses, and for distributing example sentences into senses. Furthermore, Lexonomy as a light-weight, web-based system for writing and publishing dictionaries will also support features like, for example, a mechanism for handling cross-references. In the future, users will be able to include cross-references from one entry to another entry or to a location in another entry (such as a specific sense inside another dictionary entry). Lexonomy will make sure the cross-references are clickable when the entry is formatted for display. Figure 1 shows this relationship on the left: Users interact with the Sketch Engine and Lexonomy web interfaces, and the two processes analyse their respective corpora and dictionaries. The data and functions of the other service are accessed via their API.

It should be noted that Sketch Engine is a subscription-based service, although free access for non-commercial use of Sketch Engine between 2018 and 2022 is funded[8] by the EU through ELEXIS. Lexonomy, on the other hand, is open-source software, with source code available from a GitHub repository[9] and licensed under the MIT License, which allows unrestricted re-use even for commercial purposes; so anyone can download and set up a local installation of Lexonomy and customize it to meet specific requirements. In addition, the development of Lexonomy is backed by the sponsorship of Lexical Computing (the company that makes Sketch Engine) and by funding from ELEXIS. This design provides access to the internal data representation of Lexonomy dictionaries and simplifies the task of transferring applications and data to another setup as needed; it also enables on-premise data storage, which retains the ability to

[6] https://creativecommons.org/licenses/

[7] https://opendatacommons.org/licenses/odbl/summary/

[8] https://www.sketchengine.eu/elexis/

[9] https://github.com/elexis-eu/lexonomy/

failover to a different data centre when everything else fails. Additionally, this brings about the possibility of designing one's own applications that rely on Lexonomy without much risk of a possible vendor lock-in. This is illustrated in Figure 1 on the right, where users interact with their own application, which in turn uses the API to access Lexonomy data and functionality while managing its own (private) data.

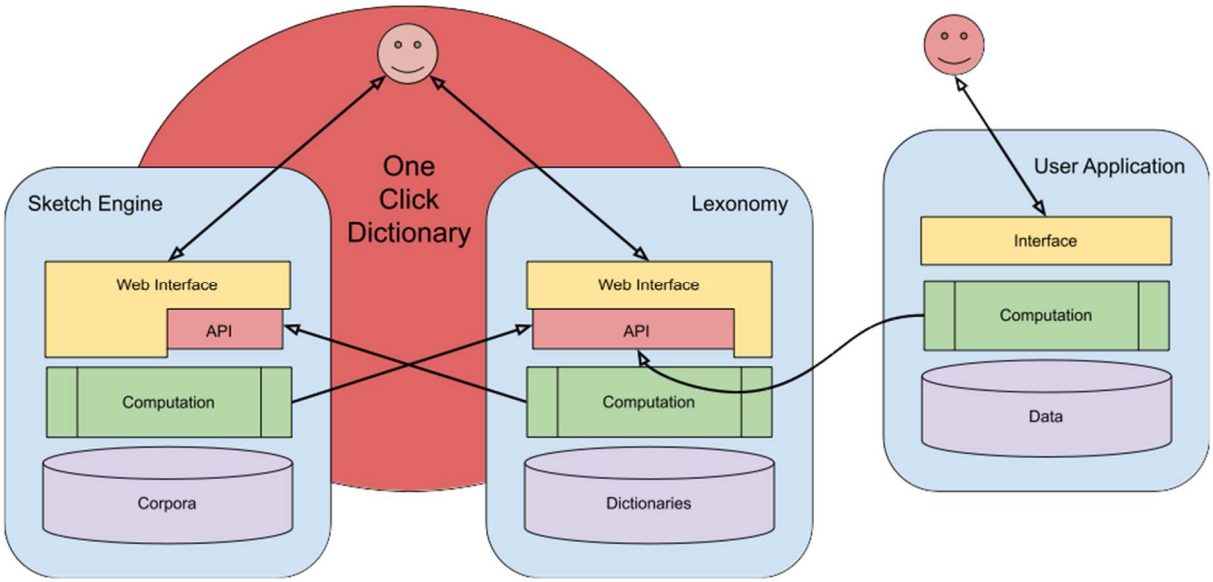

Figure 1: The One-Click Dictionary automatizes the data exchange between Sketch Engine and Lexonomy. The communication is channelled through an API, and users interact with the services via their respective web interfaces. On the other side, users can also design and use their own applications to access data in Lexonomy via an API.

Additionally, there exists another possibility: The development of a user application could also become part of Lexonomy. It is an open-source project with a growing community embedded in an ongoing European Union infrastructure project dedicated to lexicography. The users already include the University of Ljubljana, the Dutch Language Institute (Instituut voor de Nederlandse Taal), and Eurac Research (i.e. the authors of this paper). These users are also active contributors[10] to the GitHub repository, and all have participated in two previous hackathons. Both hackathons lasted approximately 2.5 days, and one was conducted with all participants on-site, the other on scheduled days with scheduled telephone and video conferencing. During these hackathons questions, problems, ideas could be discussed, joint strategies worked out and above all (partially) implemented. The general progress of the development of Lexonomy can be tracked by the contributions in the repository and the activities in the ticketing system but, above all, the development can be influenced by active participation on these channels and the dedicated Google Group[11].

[10] https://github.com/elexis-eu/lexonomy/graphs/contributors

[11] https://groups.google.com/forum/#!forum/elexis-lexonomy

For the STyrLogism Project, we have started to use Sketch Engine's web corpus capabilities, which include on-demand web crawling (also of predefined individual sites), boilerplate removal, deduplication, and tokenization, tagging, lemmatization. The boilerplate removal is applied on crawled texts to remove unwanted portions, namely navigation and menus, advertising, legal text, tabular data and any other types of text unsuitable for linguistic analysis and therefore for inclusion in a corpus. The data then undergoes a deduplication procedure where both perfect duplicates, as well as near duplicates, are removed so that only one instance of each text is preserved, and finally a Natural Language Processing pipeline divides the text into words (tokenization), enriches it with part-of-speech (PoS-tagging) and assigns the base form to each word form (lemmatization). In addition, we have begun to participate in the development of Lexonomy and advance our use-case to adapt Lexonomy as a replacement for our previous interface. We believe that the common ground between the different users will promote rich development and that we will be able to overcome certain difficulties with growing user and development communities.

4. Conclusions and Outlook

For a pending in-depth evaluation, we will use the VWB and an automatically generated "One-Click Dictionary". This will allow us to check the automatically generated lexicon, but will also allow us to put the VWB to the test with the automatically calculated data. Ideally, by using this approach, we should overcome the previously mentioned shortcomings of the VWB. So far we can at least say that a manual search for meanings of "Bar" on the latest web data—in contrast to the old newspaper data—was successful. That is, we found a use of "Bar" in the sense of "coffee shop": "In der Bar des Hotels sind auch Tagesgäste gerne willkommen und geniessen köstliche Kuchen und dazu Kaffee" ("Day guests are also welcome at the hotel's bar to enjoy delicious cakes and coffee").

Some of the pressing desiderata worth mentioning in conclusion are the availability of appropriate corpora to observe language use (including everyday situations) and detect trends of the local standard variety of STyrGerman, as well as extensive support for automatically extracting relevant data for variety lexicography (e.g. collocations, "new" word forms and meanings).

Cooperation with an international lexicographic infrastructure such as ELEXIS should strengthen the position of local varieties and dialects, provide access to current methods and tools, and also influence their design. In addition, local digital resources will be integrated into modern workflows and jointly tested.

5. References

Abel, A. (2018). Von Bars, Oberschulen und weißen Stimmzetteln: Zum Wortschatz des Standarddeutschen in Südtirol. In S. Rabanus (ed.) *Deutsch als Minderheitensprache in Italien: Theorie und Empirie kontaktinduzierten*

Sprachwandels, pp. 283–323.

Abel, A. & Stemle, E. W. (2018). On the Detection of Neologism Candidates as Basis for Language Observation and Lexicographic Endeavours: The STyrLogism Project. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pp. 535–544.

Ammon, U., Bickel, H. & Lenz, A. N. (Eds.). (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen.* 2nd, updated and extended edition. Berlin/Boston: De Gruyter Mouton.

Androutsopoulos, J. (2011). Language change and digital media: A review of conceptions and evidence. In K. Tore & N. Coupland (eds.) *Standard languages and language standards in a changing Europe*, pp. 145–161.

Barton, D. & Lee, C. (2013). *Language online: Investigating digital texts and practices.* Milton Park, Abingdon, Oxon: Routledge.

Granger, S. & Paquot, M. (eds.). (2012). *Electronic Lexicography.* Oxford, New York: Oxford University Press.

Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2011). The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In K. Allan & J. A. Robinson (eds.) *Current Methods in Historical Semantics,* pp. 59–96. https://doi.org/10.1515/9783110252903.59.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, A., Herman, O., Bušta, J., Kovář, V. & Jakubíček, M. (2015). DIACRAN: a framework for diachronic analysis. In F. Formato & A. Hardie (eds.) *Corpus Linguistics 2015: Abstract Book*. Lancaster, UK: UCREL.

Kinne, M. (1998). Der lange Weg zum Neologismenwörterbuch. Neologismus und Neologismenlexikographie im Deutschen. Zur Forschungsgeschichte und zur Terminologie, über Vorbilder und Aufgaben. In W. Teubert (ed.) *Neologie und Korpus*, pp. 63–110.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pp. 881–891.

Kupietz, M., & Lüngen, H. (2014). Recent Developments in DeReKo. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*

Měchura, M. (2017). Introducing Lexonomy: An open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference.*

O'Donovan, R. & O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In J. D. Elisenda Bernal (ed.) *Proceedings of the 13th EURALEX International Congress*, pp. 571–579.

Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisticae*, 16, 163. https://doi.org/10.14746/il.2008.16.14

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. *First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pp. 65–70.

Schäfer, R. & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani & S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Schulz, S., Lyding, V. & Nicolas, L. (2013). STirWaC: compiling a diverse corpus based on texts from the web for South Tyrolean German. In S. Evert, E. Stemle & P. Rayson (eds.) *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, pp. 35–45.

Stenetorp, P. (2010). Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus. Master's Thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.