On the Detection of Neologism Candidates as a Basis for Language Observation and Lexicographic Endeavors: the STyrLogism Project

Andrea Abel, Egon W. Stemle
Institute for Applied Linguistics, Eurac Research
E-mail: andrea.abel@eurac.edu, egon.stemle@eurac.edu

Abstract

The goal of the project STyrLogisms is to semi-automatically extract candidate neologisms (new lexemes) for the German standard variety used in South Tyrol. We use a list of manually vetted URLs from news, magazines and blog websites of South Tyrol, and regularly crawl their data, clean and process it. We compare this new data to reference corpora, additional regional word lists and all the formerly crawled data sets. Our reference corpora are DECOW14, with around 60 million word forms, and the South Tyrolean Web Corpus, with around 2.4 million word forms; the additional word lists consist 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). Here, we will report on the method employed, the first round of candidate extraction with an approach for a classification schema for the selected candidates, and some remarks on the second extraction round.

Keywords: neologism, web corpus, dictionary of variants

1 Research Goals and Motivation

The goal of the STyrLogism project is to semi-automatically extract neologism candidates for the German standard variety used in South Tyrol, a province in Northern Italy where German is an official language. 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, & Lenz 2016) and other dictionaries. More generally, the project is to 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.

The research on the German standard variety used in South Tyrol is based on the concept of pluricentricity of the German language (Clyne 1992, Ammon 1995). According to this, differing standard varieties are being used in German speaking areas. Among the crucial aspects for considering a variety a standard and not only a dialectal variety we can mention the official status of the language in a specific area, school instruction in the language, the existence of own codices, etc. South Tyrol is a particularly interesting object of linguistic studies, is due to its role as "national semi-center" (i.e. not having own language codices) from a pluricentric perspective, its marginal position within the German speaking area and the language contact situation (above all concerning the German and the Italian languages) (cf. Ammon, Bickel, & Lenz 2016). In 2016, the entirely revised second edition of the *VWB* was published 12 years after the first edition. But for this new edition it was not possible to analyze the South Tyrolean German variety to the same extent as the varieties of the "full centers" (Germany, Austria, Switzerland) and recent developments are less represented (cf. Abel 2018). We are aware that it is not among the aims of a dictionary of variants to record neologisms, but the

example shows that, in general, a constant, comprehensive language observation and documentation of the German language in South Tyrol over time is still missing, including, of course, the emergence of new words. However, the nexus between the research on language change and neologisms is evident, the former being the superordinate subject area (cf. Kinne 1998: 76).

2 Definitions of Key Concepts

Usually, investigations of lexical innovation use the following categories: neologisms, occasionalisms and other innovations. The aim of our project is the detection of neologism candidates. Neologisms are usually divided into at least the following two categories: one category for words used in a new meaning (*Neubedeutung*) but without any change of form, the other category for new lexemes (*Neulexem*) with an unseen graphical representation, for example compounds or derived forms (Kinne 1998: 83 ff.) (cf. Figure 1). According to Kinne (1998: 85) a defining feature of a neologism is that, initially, it is not included in any dictionary; it emerges from a communication need in a communication community and it passes through different phases, such as becoming common usage practice, acceptance, and lexicalization, as well as the perception of its newness from a majority of the language users.

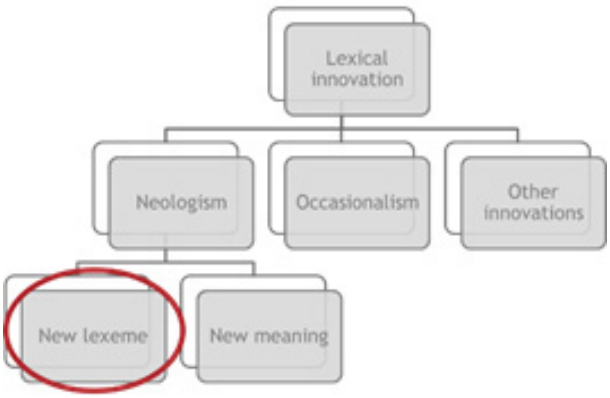

Figure 1: Lexical innovation (Kinne 1998: 86, adapted version).

In the STyrLogism project, we focus on new lexemes of the written standard local language ignoring misspellings/typos, named entities and inflected forms. Furthermore, our currently employed methodology initially focuses on the identification of neologism candidates, as only longitudinal studies with repeated observations of individual candidates can reveal true neologisms. And, as stated elsewhere, each neologism is originally an occasionalism stemming from an individual need for expression (cf. Kinne 1998: 77-78, referring to Coseriu 1958/74). As such, we are currently less interested in frequency distributions and focus on the collection of an initial data set that, consequently, may include hapax legomena as well.

The analyzed candidates are used in general language or common academic language (*alltägliche Wissenschaftssprache*, cf. Ehlich 1993). The reason to also consider commonly used technical terms in our study is due to the fact that radical social, political and economic changes activate the genesis of neologisms, necessitating designations for new circumstances, institutions etc. (cf. Kinne 1998: 87). South Tyrol was part of Austria annexed to Italy after the First World War, and thus it is obvious that a large number of neologisms are due to this radical historical change. With regard to our project, a further restriction has to be done in the sense that we are exclusively interested in STyrLogism candidates, which means in those candidates whose usage is limited to South Tyrol.

Thus, neologism candidates in the STyrLogism project can be briefly characterized as follows:

- new lexemes, not lexicalized
- used in general language or in common academic language
- consideration of the written standard language
- exclusion of misspellings/typos
- exclusion of named entities
- exclusion of inflected forms of lexicalized words
- no distinction from occasionalisms possible.

In particular, STyrLogism candidates exhibit the following features:

- neologism candidates
- usage limited to South Tyrol.

The restriction of the usage of the entities considered in the project means that they are not present in the reference corpus DECOW14 (Schäfer & Bildhauer 2012) nor in the German neologism platform *Wortwarte* (Lemnitzer 2000-2017).

3 Related Work

Up-to-date high-quality word lists and structured data is not only required for lexicography, but is also helpful for a wide range of human-language technologies (HLT), such as machine translation, named entity recognition, and spelling error detection. With the recent success of neural network methods in HLT and the related word embeddings, the need for large amounts of unlabeled data, i.e. corpora, has been emphasized, with word lists and structured data accessory parts of this. However, they are still used for supervised training to adapt to new genres, domains or languages, or for evaluation purposes. (For detailed insights into recent developments see, for example, Bethard et al. 2016; Ide, Herbelot, and Màrquez 2017; Calzolari et al. 2018). With more diachronic, genre- and domain-specific corpora becoming available, automatic neologism detection provides a head start to improve lexicographic resources and HLT tools and, as such, is becoming increasingly important.

Generally speaking, the approaches for neologism detection can be divided into two groups. One, usually applied to a single set of new data, uses language resources like word lists or linguistic patterns. The word lists are compiled from existing lexicographic resources, such as dictionaries or corpora, combined with filters for the elimination of 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 data sets, uses statistical measures or machine learning to calculate and asses the increase in usage or the change in meaning over time or in different registers. For examples, see Stenetorp (2010), Herman. and Kovár (2013) and Kilgarriff et al. (2015). Finally, these two approaches can also be combined. The STyrLogism project is currently following the former approach.

Wortwarte (Lemnitzer 2000-2017) is the most relevant previous project with regard 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 newspapers and magazines. After cleaning the HTML content, the plain text is used to build a new time slice of a corpus. The selection of appropriate neologism candidates is carried out on the basis of short-term evaluations, where the new corpus is compared with the continuously growing German reference corpus (Das Deutsche Referenzkorpus

– DeReKo. For an overview, cf. Kupietz & Lüngen 2014) with approximately 42 billion word tokens (status: 03.02.2018). To avoid "random" errors (e.g., typing errors) and filter out misspellings, the selection of neologisms is done 'manually' after the comparison with the DeReKo. The results these analyses are put online at irregular intervals, but as a rule of thumb about once a week. The results usually include a few words with their exemplary use in a sentence, and the reference as to where it came from.

O'Donovan and O'Neill (2008) use a similar idea, but in 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 is in the tradition of the Bank of English[1] rather than a static, balanced resource such as the British National Corpus (BNC). They also make use of 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 for monitoring the change 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.

4 Methods and Data

We use a list of manually selected URLs from news, magazines and blog websites of South Tyrol, and regularly crawl their data with the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler Heritrix[2]. The whole content from the crawled web pages is saved in the Web ARChive (WARC) archive format (ISO 28500), a method for combining multiple pages into an archive file together with related meta information, like retrieval date, URL, IP address. We then use Schäfer and Bildhauer's (2012) texrex toolkit for web corpus construction, which performs basic cleanups and boilerplate removal, simple connected text detection as well as shingling to remove duplicates from the corpora. The toolkit 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, i.e. 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 step removes perfect duplicates, i.e. documents that are identical up to the last character; the second step removes near-duplicates by computing token-*n*-grams for each page and the corresponding fingerprint (w-shingle). This fingerprint has the property that similar pages end up with similar fingerprints, and thus the data can easily be de-duplicated by selecting a range of allowed similarity 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) the one from our reference corpora, b) the additional word lists, which is in practice 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 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 current crawl is then tokenized – but not lemmatized – and converted into a word list. This

1 http://www.collins.co.uk/books.aspx?group=153

2 https://archive.org/projects/

list of candidate words consists of those in the current crawl that appear less than a predefined number of times in all of the other data.

Finally, the candidates are manually checked in a specifically crafted streamlined interface. This interface shows a set 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 checkbox or enter a comment into a text field (which automatically triggers the checkbox) 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 (cf. Figure 2).

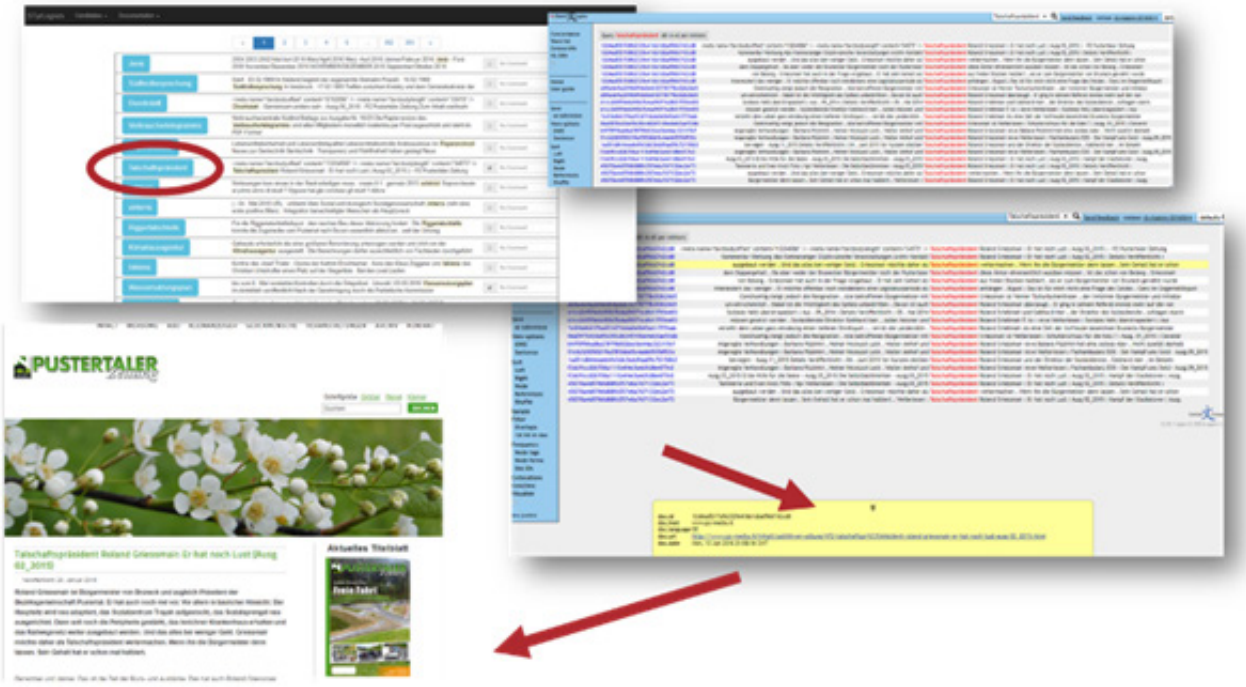

Figure 2: Web interface being used within the STyrLogism project.

5 Preliminary Results

The manually selected candidates are checked on a regular basis in order to allow a long-term monitoring. At the same time, our up-dated corpus is regularly analyzed with respect to possible new neologism candidates. Here, we will be reporting on the two rounds of manual checks of data crawled approximately two years apart from each other.

5.1 First Round: Approach for a Classification

For this round we used our initial list of 43 manually selected URLs and let the crawler run for almost two days. The minimum occurrences for a wordform to be considered were: it needed to occur at least

once in our data and was filtered out if it occurred at least once in the reference material. The result was about 70GB of raw web content from roughly 250,000 web pages. After cleaning and deduplication roughly 40,000 web pages remained. After comparing the new word forms to all our reference material roughly 4,000 neologism candidates remained.

The manual evaluation of the first extracted word list showed that many of the rejected candidates were a) two or more words written as one, i.e. the words were missing a space; b) unrecognizable words – with both a) and b) being erroneous left-overs of the boilerplate cleaning – c) foreign (mostly Italian) words, d) misspelled words and e) common words or variants of common words that are rare but established. This led to the selection of 340 candidates for further analysis. So far, the analyses of the first round of manually checked data allowed is to elaborate a preliminary classification of STyr-Logism candidates, including different kinds of emerging new word forms.

Thus, we have (a) legal and administrative common terms, e. g. *Landeszusatzvertrag* (regional amendment of a national collective agreement). Furthermore, we find (b) compounds with components of lexicalized variants of the standard German in South Tyrol recorded in the *VWB*. An example is *Optantengesetz* (a particular law for those people from South Tyrol who in 1939 opted for German citizenship and, consequently, decided to emigrate). In this case *Optant* is a lemma in the *VWB* but neither *Optantengesetz* nor other compounds are recorded as their own lemmas or as corresponding word formation units as in other cases in the *VWB*. Striking are examples such as *Luxuspensionär* (a retired person receiving a very high pension). *Pensionär* is recorded as a lemma in the *VWB* but is typically used in Switzerland, whereas in Austria and South Tyrol *Pensionist* is the commonly used term to refer to a retired person.

There are also (c) common words used in the standard German in South Tyrol which are not yet lexicalized. For this we can mention *Wahlsektion* (a part of a municipality whose inhabitants go to the same voting center). Although not a term equally used in the whole German speaking area, it is not recorded in the *VWB*. In addition, the manual checking revealed a series of (d) common words with uncommon word formation features which are at the interface between lexicon and grammar. *Mittelstandperson* (middle class person) may serve for illustration: IN this case we would expect an "-s-" as a linking element. A long-term monitoring may show if it is only a lapsus or a trend. However, we noticed several word formations following the same pattern, e. g. *Namenregelung* (naming policy). Generally speaking, there seems to be a tendency to use a linking "-s-" in compounding in South Tyrol, also similarities to its use in Austria, above all after *-g*, *-k*, *-ch* (cf. Ammon, Bickel, and Lenz 2016: LXXVI), although the picture is anything but clear (cf. Abfalterer 2007: 191).

Finally, we distinguish a category that on an interim basis we call (e) "true" neologism candidates. An illustrative example is the term *Vollautonomist* referring to a person standing up for a "full" political autonomy for South Tyrol remaining, at the same time, part of the Italian state and being, in this specific meaning, a particularity of the South Tyrolean context. It can be put up for discussion if the term is rather a new meaning than a new lexeme. However, the lexica and word lists used for our analyses did not contain the word form. A further example shows the use of an Italian loan word which is, according to Abfalterer (2007: 167ff.), one of the three main features of primary South Tyrolisms (i.e. variants which are supposed to occur only in South Tyrol) next to loan translations and "others". In the compound *Vollkornpizzetta* (small pizza made of whole grain) the Italian *pizzetta* with the diminutive suffix *-etta* is used. It is still debatable where to draw the line between (c) and (e), as the mentioned forms are commonly known.

Given that the category of "true" neologism candidates is particularly relevant within our study, an attempt to carry out a preliminary characterization was done. With regard to the goals of this category, different key aspects became apparent. Thus, the lexical items are used for (1) humorous, ironic

or sarcastic and, furthermore, for (2) polemic or malicious ways of expression. (3) Creative language usage and the play on words can be observed as well, and this may also appear together with (1) and (2). *Donnerwetterer* is a case in point designating in an original way someone railing against someone/something; *Donnerwetter* is a common German word, originally mainly used to refer to a thunderstorm, but nowadays referring to a loud confrontation or used as an interjection to express either anger or admiring astonishment; however, the unit, with the suffix *-er* that in German word formation is typically used to refer to persons, has not been lexicalized. Finally, the (4) designation of new circumstances, facts and objects can, of course, be found. for example, *Bausündennachlass* is used to indicate a legal measure in Italy for remitting a financial penalty for an eyesore.

The items have a number of typical features. With respect to word formation we notice a tendency towards (i) compounding (cf. also Abfalterer 2007: 189), and partly complex compounds are to be found. For instance, if we take *Regiokornbrot* then *Regiokorn* is used to designate regional grain, originally deriving from the name of a local project with the title *Regiokorn*; subsequently the term was being used more generally for bread made of local grain. Closely connected to this phenomenon is the (ii) strategy of turning phrasemes into single words. In the case of *Mundaufreißer* the idiomatic expression *das Maul* (also: *Mund*) *aufreißen* (to give oneself airs, literally to open the mouth wide (*Maul* in German regards to an animal, *Mund* to a person)) is used in an unusual way as a compound that is conflicting with the principle of fixedness of phraseologisms (cf. Burger 2007).

As expected, some candidates constitute (iii) loan words or loan translations from the Italian language. Here we might mention *Promotorenkomitee* (a committee of initiators, supporters of an action) which is a literal translation of the Italian *comitato promotore*, commonly used in South Tyrol, whereas in other German speaking areas words such as *Initiatoren* or *Befürworter* are used instead. We also have (iv) formal analogies to lexicalized variants. An interesting case is *Schwammlklauber* (a person picking mushrooms), a commonly used word form in South Tyrol but not recorded in the *VWB*. However, *Schwammerl* (mushroom) is a lemma in the *VWB* (used – with the diminutive suffix *-erl* – in southeast Germany and in Austria which, according the rules applied for the *VWB*, means that the usage in South Tyrol is implied, cf. Ammon, Bickel, and Lenz 2016: LXXVI) but not the form *Schwamml*, which is the typical word form in the South Tyrolean context (the assumption of the use of the suffix *-erl* also in South Tyrol in the *VWB* is shown in other cases as well, e.g. concerning the lemma *Sackerl*, i.e. a carrying bag, which is not used in South Tyrol). The verb *klauben* (to harvest, to pick) is also lexicalized in the *VWB* (used in southeast Germany and in Austria) containing the diasystematic label “borderline case of the standard language”. On the other hand *Apfelklauber* is recorded as primary South Tyrolism in the *VWB*, and this without any diasystematic label. However, it has an own, limited meaning as it refers to a person helping to harvest apples being paid for this activity, whereas *Schwammlklauber* indicates someone doing the activity for leisure.

Among the affected domains, it is worth mentioning politics, environment, tourism, leisure and food.

5.2 Second Round: Some Remarks

For this round we used an updated list with 156 manually selected URLs and let the crawler run for three days. The minimum occurrences for a wordform to be considered were: it needed to occur once in our data and was filtered out if it occurred once in the reference material. The result was about 60GB of raw web content from roughly 500,000 web pages. After cleaning and deduplication roughly 50,000 web pages remained. After comparing the new word forms to all our reference material, roughly 7,000 neologism candidates remained. From the monitored candidates, only seven reappeared in the new data set.

Although the overlapping of the comparison was low, we might have a closer look at one of the word fields affected. It is notable that morphological variations of *autonomiefreundlich* (autonomy-friendly) and *autonomiefeindlich* (anti-autonomy) reappeared in the second round. These lexical units form a part of a word field of a constantly hot topic in the South Tyrolean context, being the political autonomy perceived as an important achievement for the German speaking population (cf. Autonome Provinz Bozen Südtirol: 2004). Thus, we also found *Vollautonomist* in the first round. Furthermore, a data checking in the DECOW14 corpus (Schäfer & Bildhauer 2012) confirmed the former usage of *Vollautonomie* ("full" autonomy) which, to the best of our knowledge, has never been in discussion for inclusion in the *VWB*. Furthermore, we can find the commonly used *Autonomie* (autonomy) exclusively in this narrow political sense, including patterns such as *dynamische Autonomie* (dynamic autonomy).

6 Conclusions and Outlook

In the paper we gave an overview on our work focusing on those neologism candidates with the potential to persist over time and to be lexicalized. The first findings of the initiative show that the approach is suitable to produce candidate lists for newly arisen words, or rather word forms not included in the corpora and word lists to date, even though a large amount of noise had to be eliminated manually. Within the time period taken into consideration and with the data basis used so far it was not possible to distill a larger amount of lexical units being characterized by persistence over time. However, the approach turned out to be a useful support for the overarching endeavor of language observation and documentation in South Tyrol.

We found that the online publishing attitude in South Tyrol makes our task more difficult in two ways: first, major newspaper and magazine publishers in the region often only put an excerpt or summary of an article online. This reduces the amount of actual text that can be used for our analyses, and also complicates the extraction of content from single web pages: extracting content from a web page is a balancing act between getting as much of the desired textual content as possible (recall), but at the same time *only* getting the desired content and not the superfluous boilerplate (precision). This task generally gets more difficult with lower amounts of available content, and produces more noise with a lower content-to-boilerplate-ratio (Schäfer & Bildhauer 2012). Second, articles only stay online for a short period of time. This period, depending on unknown factors, can be as short seven days.

Consequently, a methodological possible next step includes to shorten our crawl interval to be around the minimum content availability time in the region. This would mitigate the otherwise unavoidable loss of early onsets of new word forms and, additionally, would also enable precise time series analyses for word usage over time. To this end, we could use the SketchEngine's "Trends: Neologisms and diachronic analysis of word usage" feature (cf. Herman & Kovár (2013) for the version currently implemented in the SketchEngine) as a start and see whether this yields promising results. Later, we could adapt the idea for our particular use-case.

Utilizing social media and thereby extending the basis for the data analyses could also prove helpful: users produce a tremendous amount of text each day on social media, much of which is readily available without the complications of boilerplate removal, as needed for web pages. This development has opened new possibilities for lexicographical analyses, such as, in "particular, corpus patterns that are very rare in conventional-size corpora turn out to have many occurrences in the very large corpora of social media" (Cook 2012).

A different direction could also be to detect novel senses, i.e. semantic changes in established word forms, based on distributional similarity between word models built from different corpora (cf.

Gulordava & Baroni (2011) for a successful application of vector space models in this context). Here, we could also employ word embeddings (Mikolov et al. 2013), a recently very successful language modeling technique with results, often on par or superior to the established vector space models.

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*. - Germanistische Linguistik: Themenheft, pp. 283-323

Abfalterer, H. (2007): *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht*. Innsbruck: Innsbruck University Press.

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

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 ed. Berlin/Boston: De Gruyter Mouton.

Autonome Provinz Bozen Südtirol (eds.) (2004): Südtirol-Handbuch. 23th ed. Bolzano/Bozen: Landespresseamt.

Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P. & Zesch, T. (eds.) (2016). *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics.

Burger, H. (2007). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

Ehlich, K. (1993). Deutsch als fremde Wissenschaftssprache. In A. Wierlacher et al. (eds.) *Jahrbuch Deutsch als Fremdsprache*, 19. München: iudicium, pp. 13-42.

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*. Tübingen: Gunter Narr Verlag, pp. 63-110.

Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. & Tokunaga, T. (eds.) (2018). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).

Cook, P. (2012). Using social media to find English lexical blends. In R. V. Fjeld, J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 846–854.

Gulordava, K. & Baroni, M. (2011). A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 67–71.

Herman, O. & Kovár, V. (2013). Methods for Detection of Word Usage over Time. In *Proceedings of the Seventh Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2013)*. Brno, Czech Republic: Tribun EU

Ide, N., Herbelot, A. & Màrquez, L. (eds.) (2017). *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Association for Computational Linguistics.

International Organization for Standardization (2017). Information and documentation - WARC file format (ISO 28500).

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*. Berlin/Boston: De Gruyter, pp. 59–96. http://doi.org/10.1515/9783110252903.59

Kilgarriff, A., Ondřej, H.. Bušta, J., Rychlý, P. & Jakubíček, M. (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics (CL2015)*, United Kingdom.

Kupietz, M. & Lüngen, H. (2014). Recent Developments in DeReKo. In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Lemnitzer, L. (2000-2017). Die Wortwarte. Accessed at: http://wortwarte.de/ [April 28, 2017]

Mikolov, T., Corrado, G., Chen, K. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12.

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. E. Bernal (ed.) *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 571–579.

Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. In *Investigationes Linguisticae, XVI*; Adam Mickiewicz University: Poznań, Poland.

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

Schäfer, R. & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In N. Calzolari et al. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

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)*. Lancaster, UK, pp. 35–45.

Stenetorp, P. (2010). Automated extraction of swedish neologisms using a temporally annotated corpus. Stockholm, Sweeden: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.