

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

Andrea Abel & Egon Stemle  
Institute for Applied Linguistics  
Eurac Research, Bolzano/Bozen, Italy

[andrea.abel@eurac.edu](mailto:andrea.abel@eurac.edu)

[egon.stemle@eurac.edu](mailto:egon.stemle@eurac.edu)

# Overview

- Research goals
- Some definitions
- Related work
- Method
- Data
- Preliminary results
- Conclusions & outlook

## Research goals

- Semi-automatic extraction of neologism candidates for the German standard variety used in South Tyrol (Northern Italy)



- Language observation and evaluation of trends of the local standard variety of the German language
- Consideration for future editions of the “Variantenwörterbuch des Deutschen” (Dictionary of variants of the German language) (Ammon et al. 2016) and other dictionaries

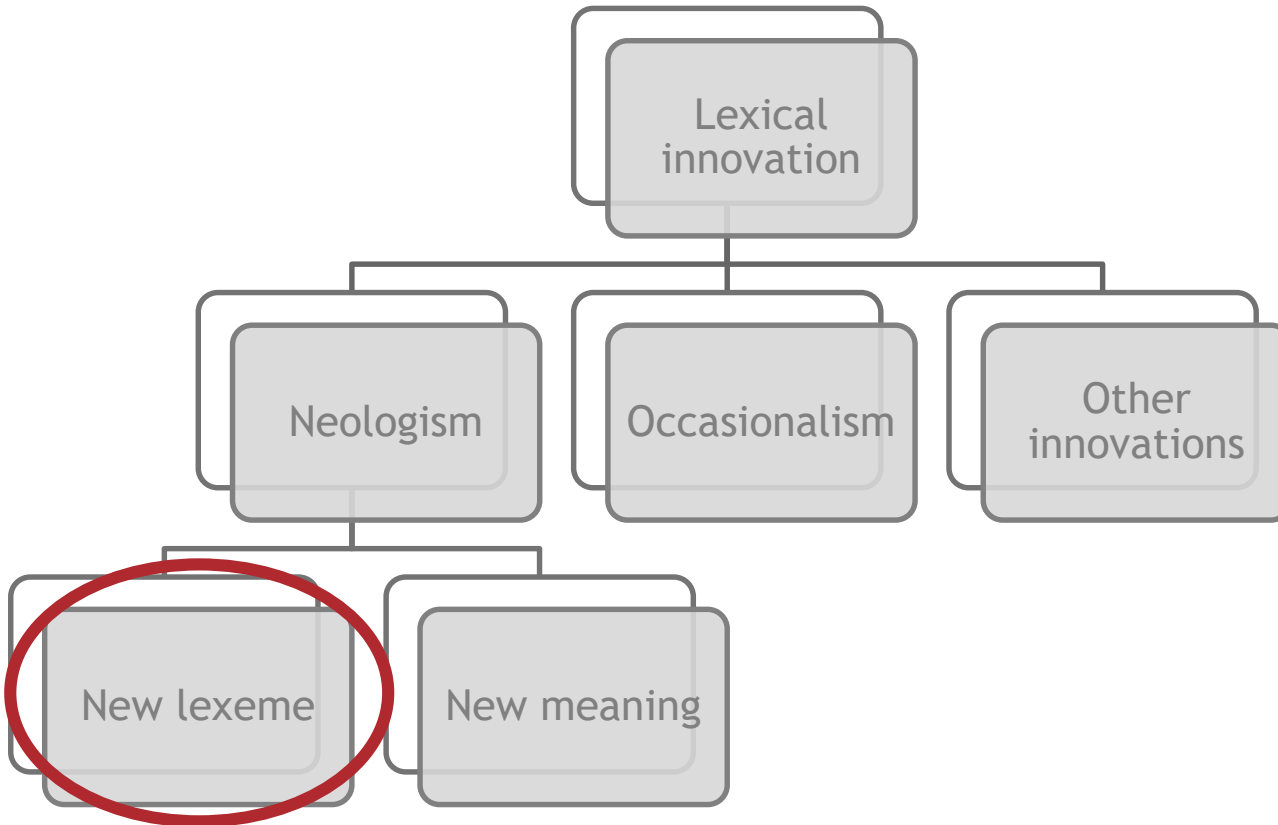
## A small excursus ...

- Research on the German standard variety used in South Tyrol



- Based on the concept of pluricentricity of the German language (cf. Clyne 1992, Ammon 1995):  
differing standard varieties used in the German speaking area (official status, taught in school, codices, etc.)
- South Tyrol as an interesting object of linguistic studies:
  - role as “national semi-centre” from a pluricentric perspective
  - marginal position within the German speaking area
  - the language contact situation

## Some definitions



(Kinne 1998: 56, adapted version)

# Some definitions

Neologism candidates:

- new lexems, not lexicalised
- used in general language or common academic language (“alltägliche Wissenschaftssprache”, cf. Ehlich 1993, 1999)
- consideration of the written standard language
- no misspellings/typos
- no named entities
- no inflected forms of lexicalised words
- no distinction from occasionalisms possible

## Some definitions

STyrLogism candidates:

- neologism candidates
- usage limited to South Tyrol  
(not present in the German reference corpus DECOW14, Schäfer/Bildhauer 2012, and in the German neologism platform “Wortwarte”, Lemnitzer 2000-2017)

(Remark: Frequency currently not taken into consideration)

## Related work

Different approaches for neologism detection:

1. Use language resources, like *known words* or *word patterns* (approach is often applied to a single set of new data)
2. Use statistical measures or unsupervised machine learning (approach is often applied to multiple data sets, e.g. diachronic data)
3. Use a combination of 1. and 2.





# Related work

## 1. Use of language resources:

- Methods based on *known words* use word lists compiled from existing lexicographic resources, such as dictionaries or corpora, combined with filters for the elimination of non-words, typographical errors, named entities, etc.
- Methods based on *word patterns* use lexical cues, e.g. markers of lexical novelty like punctuation marks that can signal new words.

(O'Donovan 2008, Paryzek 2008)

## Related work

### 2. Use of statistics and (unsupervised) machine learning:

- Statistical measures and machine learning methods can be applied to calculate and assess an increase in usage over time.

(Kilgarriff 2015, Stenertorp 2010)

### 3. Combination of methods:

- These methods can also be combined, like training a classifier to recognise successful and promising word formation patterns, e.g. on the basis of manually classified unknown words.

# Method

1. Retrieve web pages (of pre-selected web sites) through web crawling
2. Clean the web pages (Boilerplate removal, deduplication, etc.)
3. Build a corpus, extract word list, filter with *known words* and remove non-words
4. Present list of possible neologisms to linguist(s) for analysis

This is similar to e.g.

- NeoCrawler (Kerremans et al. 2012)
- Wortwarte (Lemnitzer 2000 - 2017)
- Neoloog (being developed at INT; Stemle / Jakubiček / Tiberius ENeL 2015 presentation)

# Data (1st round)

## Data preprocessing:

- 44 seed URLs (online newspapers, magazines) used for web crawling in June 2016
- After 2 days of crawling 250k URLs were processed
- After removing very short and very long documents, and deduplication 40k documents were left
- After paragraph deduplication the final size of the corpus was: 11Mio tokens (200k unlemmatised types)

# Data (1st round)

Reference Data (60M + 30k types):

- General German Web Corpus (12Bio tokens with 60Mio types): DECOW14
- Named entities, terminological terms and other lists (~30,000 types):
  - AU-CH-STyr-ismen: 3,272 type
  - STyr NEs: 12,499 types
  - VWB: 14,466 types
  - Wortwarte

After comparing the data sets ~4,500 STyrLogism candidates were left for manual checking.

## Data (2nd round)

### Data preprocessing:

- 156 seed URLs (online newspapers, magazines) used for web crawling in March 2018
- After 3 days of crawling 500k URLs were processed
- After removing very short and very long documents, and deduplication 54k documents were left
- After paragraph deduplication the final size of the corpus was: 35Mio tokens (260k unlemmatised types)

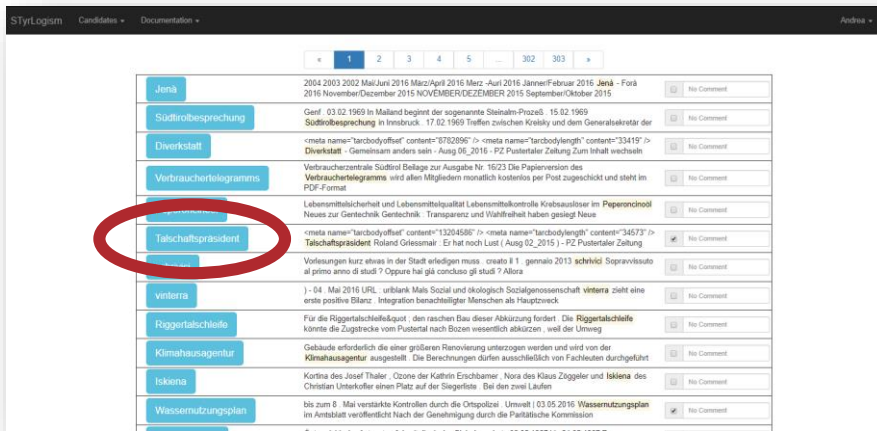
## Data (2nd round)

Reference Data (60M + 30k types):

- General German Web Corpus (12Bio tokens with 60Mio types): DECOW14
- Named entities, terminological terms and other lists (~30,000 types):
  - AU-CH-STyr-ismen: 3,272 type
  - STyr NEs: 12,499 types
  - VWB: 14,466 types
  - Wortwarte

After comparing the data sets **~7,600** STyrLogism candidates were left for manual checking.

# Web interface





# Web interface

The screenshot displays a web interface with a left sidebar containing a list of documents. The 'Tatschatspräsident' entry is circled in red. The right pane shows search results for 'Tatschatspräsident' with a red arrow pointing from the circled entry to the search results.

Document Title	Date	Comments
Jenä	2004 2003 2002 Mai/Juni 2016 März/April 2016 März-April 2016 Januar/Februar 2016 Jenä - Forä 2016 November/December 2015 NOVEMBER/DECEMBER 2015 September/Oktober 2015	No Comment
Südtirolbesprechung	Genf 03.02.1960 in Mailand beginnt der sogenannte Südtirol-Prozess 15.02.1960 Südtirolbesprechung in Innsbruck 17.02.1960 Treffen zwischen Kriegerly und dem Generalsekretär der	No Comment
Diversität	<meta name="tacobodyoffset" content="3782086" /> <meta name="tacobodylength" content="33415" /> Diversität - Gemeinsam anders sein - Ausg 06_2016 - PZ Pustertaler Zeitung Zum Inhalt wechseln	No Comment
Verbraucherleitgramms	Verbraucherzentrale Südtirol Beilage zur Ausgabe Nr. 16/23 Die Papierversion des Verbraucherleitgramms wird allen Mitgliedern roudentlich kostenlos per Post zugeschickt und steht im PDF-Format	No Comment
Tatschatspräsident	Lebensmittelsicherheit und Lebensmittelsicherheit Lebensmittelsicherheit Krebsauslöser im Pappenschmel Neues zur Gentechnik Gentechnik: Transparenz und Wahlrecht haben gesiegt Neue <meta name="tacobodyoffset" content="13204680" /> <meta name="tacobodylength" content="34573" /> Tatschatspräsident Roland Griesmayr: Er hat noch Lust (Aug 02_2015) - PZ Pustertaler Zeitung	No Comment
Vinterra	Vorlesungen kurz etwas in der Stadt erledigen muss. creato 8 1 gennaio 2013 scrividi Sopravvissuto al primo anno di studi? Oppure hai già concluso gli studi? Allora	No Comment
Riggertalschleife	14. Mai 2016 LURL - uribank Mals Sozial und ökologisch Sozialgenossenschaft vinterra zieht eine erste positive Bilanz: Integration benachteiligter Menschen als Hauptzweck Für die Riggertalschleife könnte die Zugetrücke vom Pustertal nach Bozen wesentlich abkürzen, weil der Umweg	No Comment
Klimahausagentur	Gebäude erforderlich die einer größeren Renovierung unterzogen werden und wird von der Klimahausagentur aufgestellt. Die Berechnungen dürfen ausschließlich von Fachleuten durchgeführt	No Comment
Wassernutzungsplan	Kortina des Josef Thaler, Ozone der Katrin Erschbamer, Nora des Klaus Zoggele und Isabella des Christian Unterkofler einen Platz auf der Siegerliste. Bei den zwei Läufen bis zum 8. Mai verankerte Kontrollen durch die Ortopolizei. Umwelt   03.05.2016 Wassernutzungsplan Im Anleitab veröffentlicht nach der Genehmigung durch die Paritätische Kommission	No Comment

# Web interface

The screenshot displays a search engine interface with a search bar at the top containing the query 'Tatschspräsident'. Below the search bar, there are two main panels. The left panel shows a list of search results, with the top result titled 'Tatschspräsident' circled in red. The right panel shows the full text of the selected result, which is a news article about the election of Roland Griessmair as the president of the Pustertal region. The article text is highlighted in yellow. At the bottom of the page, there are two red arrows pointing from the search results and the article text to a small box containing metadata for the document, including its ID, language, and URL.

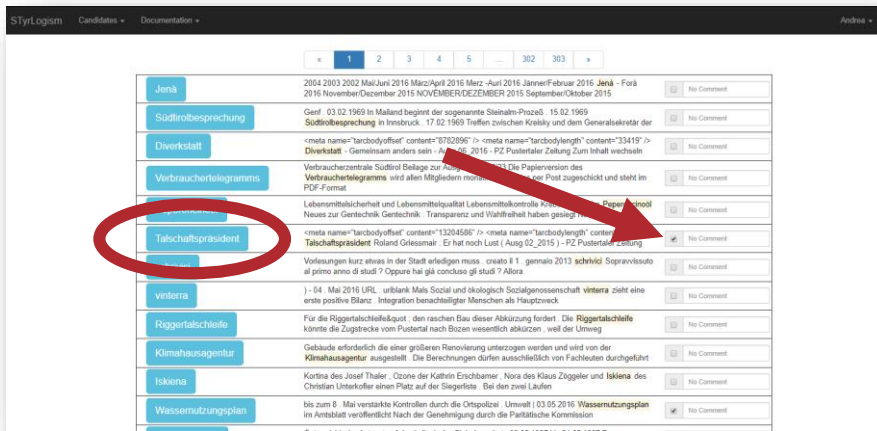
**Tatschspräsident Roland Griessmair: Er hat noch Lust (Auszug 02\_2015)**

Veröffentlichung: 29. Januar 2015

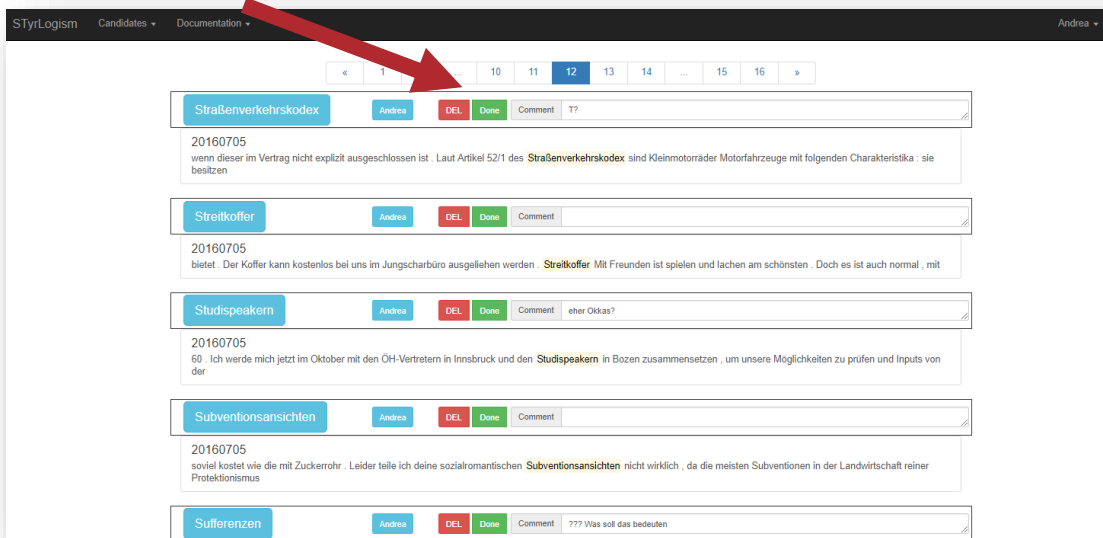
Roland Griessmair ist Bürgermeister von Bruneck und zugleich Präsident der Bezirksgemeinschaft Pustertal. Er hat noch nicht viel vor. Vor allem in beruflicher Hinsicht: Der Hauptzitz wird neu adaptiert, das Sozialzentrum Trayah aufgestockt, das Sozialspargel neu ausgerichtet. Dann soll noch die Peripherie gestärkt, das Innicher Krankenhaus erhalten und das Radwegenetz weiter ausgebaut werden. Und das alles bei weniger Geld. Griessmair möchte daher als Tatschspräsident weitermachen. Wenn ihn die Bürgermeister denn lassen. Sein Gehalt hat er schon mal halbiert.



# Web interface



# Web interface



# Preliminary results

First round:

- list of 43 manually selected URLs
- cleaning & comparison with reference material
- manual evaluation of the ~ 4,000 remaining STyrLogism candidates
- selection of 340 candidates for further analysis

# Preliminary results

Attempt to a preliminary classification of STyrLogism candidates:

„Landeszusatzvertrag“  
(regional amendment of a national collective agreement)

T (Terms):  
legal & administrative common terms

K (Compounds):  
compounds with components of lexicalised variants of the standard German (STIR)

„Luxuspensionär“  
(a retired person receiving a very high pension)

„Wahlsektion“  
(a part of a municipality whose inhabitants go to the same voting center)

V (Variants):  
common words used in the standard German (STIR) but not yet lexicalised

M (Morphological features):  
common words with uncommon word formation features

„Mittelstandperson“  
(middle class person)

N (Neologisms):  
„true“ neologism candidates

„Vollautonomist“  
(person standing for a „full“ political autonomy remaining part of the Italian state)

# Preliminary results

Second round:

- list of 156 manually selected URLs
- cleaning & comparison with reference material
- list of ~7,600 STyrLogism candidates
- reappearance of only 7 monitored candidates from the first round
- ! manual evaluation still outstanding !

# Preliminary results

Exemplary word field: “autonomy” (in a political sense)

- second round: morphological variations of “autonomiefreundlich” (autonomy-friendly) and “autonomiefeindlich” (anti-autonomy)
- first round: “Vollautonomist”
- similar forms present in reference data, e. g. “dynamische Autonomie” (dynamic autonomy)



# Conclusions & outlook

## First findings:

- suitable approach for finding word forms not included in the reference material
- not yet possible to distil a larger amount of lexical units persisting over time

## Challenges & next steps:

- content shared via few media outlets
- paywall for larger text snippets
- use of CMC data

Software available under an open source license (ASF 2.0)

<https://gitlab.inf.unibz.it/commul/styrlogism/>

# Thank you for your attention!



Andrea Abel  
([andrea.abel@eurac.edu](mailto:andrea.abel@eurac.edu))

&

Egon Stemle  
([egon.stemle@eurac.edu](mailto:egon.stemle@eurac.edu))



Institute for Applied Linguistics

([www.eurac.edu/iscm](http://www.eurac.edu/iscm))

# 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. *ional Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

# References

- Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odiijk, 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. & Jakubiček, M. (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics (CL2015)*, United Kingdom.

# References

- Kupietz, M. & Lungen, 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 Lingüística 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, Sweden: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.
- Andrea Abel, Egon Stemle