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
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

  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

Lexical innovation

Neologism

Occasionalism

Other innovations

New lexeme

New meaning

(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, Stenetorp 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
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:**

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T (Terms):</strong> legal &amp; administrative common terms</td>
<td>„Landeszusatzvertrag“ (regional amendment of a national collective agreement)</td>
</tr>
<tr>
<td><strong>K (Compounds):</strong> compounds with components of lexicalised variants of the standard German (STIR)</td>
<td>„Luxuspensionär“ (a retired person receiving a very high pension)</td>
</tr>
<tr>
<td><strong>V (Variants):</strong> common words used in the standard German (STIR) but not yet lexicalised</td>
<td>„Wahlsektion“ (a part of a municipality whose inhabitants go to the same voting center)</td>
</tr>
<tr>
<td><strong>M (Morphological features):</strong> common words with uncommon word formation features</td>
<td>„Mittelstandperson“ (middle class person)</td>
</tr>
<tr>
<td><strong>N (Neologisms):</strong> „true“ neologism candidates</td>
<td>„Vollautonomist“ (person standing for a „full“ political autonomy remaining part of the Italian state)</td>
</tr>
</tbody>
</table>
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)

&

Egon Stemle
(egon.stemle@eurac.edu)

Institute for Applied Linguistics
(www.eurac.edu/iscm)
References


References


References


