News from the STyrLogism Project:
Language varieties meet
One-Click Dictionary
Korpuslinguistische Ermittlung und
lexikographische Bearbeitung
Workshop am IDS, Mannheim

egon w. stemle
<egon.stemle@eurac.edu>
Who are we?
In South Tyrol multilingualism involves geographic, institutional, social and personal aspects. The main aim of our research is to provide scientific answers to current issues of language and education policy as well as to economic and social questions at local and international level.
What are we doing?
What are we doing?

4th LEARNER CORPUS RESEARCH CONFERENCE

LCR 2017

Bolzano/Bozen, 5-7 October 2017

The 4th Learner Corpus Research Conference – LCR 2017 – will take place from 5-7 October 2017 in Bolzano/Bozen in the northern Italian province of Bolzano – Alto Adige / Bozen – Südtirol. As a multilingual province, Bolzano/Bozen provides perfect surrounding conditions for a conference on learner language. LCR 2017 will be hosted by the Institute for Applied Linguistic of EURAC Research. The theme of LCR 2017 is “Widening the Scope of Learner Corpus Research”. The keynote speakers of LCR 2017 will address the theme in their lectures on L1 writing development and Learner Corpus Research (Philip Durrant), quantitative methods in Learner Corpus Research (Stefan Th. Gries), and Learner Corpus Research and Italian as L2 (Stefania Spina).

The LCR conferences, organized under the aegis of the Learner Corpus Association, bring together researchers and language teachers, software developers and others interested in Learner Corpus Research. The LCR 2017 conference aims to be a showcase for the latest developments in the field of Learner Corpus Research and will feature keynote lectures, parallel sessions for the presentation of scientific results and work in progress reports, poster presentations, software demos, a book exhibition, as well as social events for participants. There will also be a competition for best paper presentation and best poster presentation by a PhD student.

The Call for Papers has been published on 30 September 2016. Please click on the button below to see the full CFP.
What are we doing?

CMC-Corpora 2017 @ Eurac Research
October, 3-4th 2017
Research goals

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

- 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
Lexical innovation

- Neologism
  - New lexeme
  - New meaning

- Occasionalism

- Other innovations
STyrLogism

Neologism candidates

- new lexems, not lexicalised
- used in general language or common academic language ("alltägliche Wissenschaftssprache", see 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

STyrLogism candidates

- neologism candidates
- usage limited to South Tyrol
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.
**Approach for STyrLogism detection**

- Retrieve web pages (of approx. 150 pre-selected web sites) through web crawling
- Clean the web pages (Boilerplate removal, deduplication, etc.)
- Build a corpus, extract word list, filter with known words and remove non-words
- Present list of possible neologisms to linguist(s) for analysis

This is similar to, for example:

- NeoCrawler (Kerremans et al. 2012)
- Wortwarte (Lemnitzer 2000 - 2017)
- Neoloog (being developed at INT; Stemle /Jakubiček / Tiberius ENeL 2015 presentation)
Reference data for STyrLogism detection (60M + 30k types)

- General German Web Corpus (12Bio tokens with 60Mio types):
  - DECO14

- NEs, terminological terms and other lists (approx. 30,000 types):
  - AU-CH-STyr-ismen: 3,272 types
  - STyr NEs: 12,499 types
  - VWB: 14,466 types
  - Wortwarte
STyrLogism

A 2nd round of data processing

- list of 156 manually selected URLs
- cleaning & comparison with reference material
- list of approx. 7,600 STyrLogism candidates
- reappearance of only 7 monitored candidates from the 1st round
STyrLogism – Web Interface
<table>
<thead>
<tr>
<th>T (Terms)</th>
<th>K (Compounds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>legal &amp; administrative common terms</td>
<td>compounds with components of lexicalised variants</td>
</tr>
<tr>
<td>Landeszusatzvertrag</td>
<td>Luxuspensionär</td>
</tr>
<tr>
<td>(regional amendment of a national collective agreement)</td>
<td>(a retired person receiving a very high pension)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V (Variants)</th>
<th>M (Morphological features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>common words used in the standard German (STIR) but not yet lexicalised</td>
<td>common words with uncommon word formation features</td>
</tr>
<tr>
<td>Wahlsektion</td>
<td>Mittelstandperson</td>
</tr>
<tr>
<td>(a part of a municipality whose inhabitants go to the same voting center)</td>
<td>(middle class person)</td>
</tr>
</tbody>
</table>
Neologisms

ture candidates

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

more examples

autonomiefeindlicher
autonomiefreundlich
koinvolvierten
tagverträumt
Überschreibekosten
It is an unfortunate reality that academia does not consider software engineering or tool-building valid areas of academic work—but if libraries like ggplot2 and scikit-learn were referenced as often as they were used in research papers, that paradigm might shift.

Please cite.
On Tools and Standards

How standards proliferate:

(See: A/C chargers, character encodings, instant messaging, etc)

**Situation:**
There are 14 competing standards.

14?! Ridiculous!
We need to develop one universal standard that covers everyone's use cases.
Yeah!

**Soon:**

**Situation:**
There are 15 competing standards.
The bugs look so happy!
STyrLogism - in the Sketch Engine

The image shows the dashboard of the Sketch Engine with the focus on the STyrLogism corpus. The dashboard includes options for managing the corpus and accessing various tools such as Word Sketch, Word Sketch Difference, Thesaurus, Concordance, Parallel Concordance, Wordlist, N-grams, Keywords, Trends, and One-Click Dictionary. Recent results and favourite results are also displayed. The content is related to corpus searching and corpus building, with a mention of a boot camp for 2 days of corpus searching and corpus building, offering details and registration.
Welcome to Lexonomy, a cloud-based, open-source platform for writing and publishing dictionaries.

**Gentle introduction to Lexonomy**
A short and sweet guided tour of Lexonomy for beginners.

**Introducing Lexonomy: an open-source dictionary writing and publishing system PDF**
A conference paper offering a detailed review of Lexonomy’s features.

**Reference**

If you are referring to Lexonomy from an academic publication, it is recommended that you cite this paper.

Lexonomy is developed as part of ELEXIS project.
STyrLogism - Lexonomy on github

A cloud-based, open-source system for writing and publishing dictionaries. [http://www.lexonomy.eu/](http://www.lexonomy.eu/)

- lexicography
- dictionary-software
- dictionary-writing-system

- 336 commits
- 2 branches
- 0 releases
- 6 contributors
- MIT License

- mjakubicek: Python backend runs a profiler if "prof" is given in query string
- Latest commit signed 17 days ago


Sarah Schulz, Verena Lyding, and Lionel Nicolas. “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)*. Ed. by Stefan Evert, Egon Stemle, and Paul Rayson. Lancaster, UK, 2013, pp. 35–45.