NEW GERMAN WORDS: DETECTION AND DESCRIPTION

XVIII. EURALEX International Congress, Ljubljana, 17.-21. July 2018

Annette Klosa-Kückelhaus and Harald Lüngen
OUTLINE

1. The “Neologismenwörterbuch” at IDS Mannheim
2. Short information on related work
3. Finding Neologism Candidates
4. Evaluation of the Quantitative Method
5. Outlook
1. THE "NEOLOGISMENWÖRTERBUCH"

Some facts

- the “Neologismenwörterbuch” covers new words and new meanings established in the past thirty years (1991-2000, 2001-2010, 2011-today)
- continuous addition of new entries
- online publication in the dictionary portal OWID (www.owid.de) at IDS Mannheim

http://www.owid.de/docs.neo/wortartikel.jsp
1. THE "NEOLOGISMENWÖRTERBUCH"

The “Neologismenwörterbuch” comprises

- single word entries (e.g., Avatar)
- multi-word expressions (e.g., in der Pipeline)
- new elements of word formation (e.g., [...]holic)
- new meanings (e.g. texten ‘send a [short] test message in electronic media’)

http://www.owid.de/artikel/401238?module=neo&pos=13
# 1. THE "NEOLOGISMENWÖRTERBUCH"

<table>
<thead>
<tr>
<th>All entries</th>
<th>almost 1.900</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neologisms from 1991-2000</td>
<td>over 1,000</td>
</tr>
<tr>
<td>Neologisms from 2001-2010</td>
<td>almost 700</td>
</tr>
<tr>
<td>Neologisms since 2011</td>
<td>almost 150</td>
</tr>
<tr>
<td>New lexemes</td>
<td>almost 1,550</td>
</tr>
<tr>
<td>New elements of word formation</td>
<td>almost 20</td>
</tr>
<tr>
<td>New meanings</td>
<td>over 160</td>
</tr>
<tr>
<td>New multi-word units</td>
<td>almost 120</td>
</tr>
<tr>
<td>Other new lexemes (synonyms, other sense-related words, derivations, compounds, etc.) contained in entries and accessible via list</td>
<td>almost 5,000</td>
</tr>
</tbody>
</table>
1. THE "NEOLOGISMENWÖRTERBUCH"

Lexicographic information
- etymology, orthography, pronunciation
- meaning and usage
- grammar, word formation (products)
- encyclopedic information, illustrations
- frequency development

A. Klosa-Kükelhaus/H. Lüngen – „New German Words: Detection and Description“– Ljubljana, EURALEX 2018
1. THE "NEOLOGISMENWÖRTERBUCH"

**Definition**

A neologism is a lexical unit or a meaning which

- emerges in a communication community in a specific period of time of language development,
- which diffuses,
- is generally accepted as language norm,
- and which the majority of speakers perceives as new for some time.

All neologisms in the dictionary are

- fully lexicalized lexemes
- not nonce words.

A. Klosa-Kückelhaus/H. Lüngen – „New German Words: Detection and Description“— Ljubljana, EURALEX 2018
2. RELATED WORK

Uwe Quasthoff: Deutsches Neologismenwörterbuch (2007)
- strictly corpus-driven
- almost 2,300 entries
- definition of „neologism“: word whose frequency has increased significantly between 2000 and 2006
- comprises fully lexicalized lexemes and nonce words

Corpus Linguistic Approach:
- corpora partitioned into sub-corpora by year
- identification of typical frequency timelines of neologisms with a rise of frequency in the present and with minimum frequency conditions
2. RELATED WORK

Lothar Lemnitzer: Die Wortwarte (2000-today)

• automatically extracted from web corpus
• aims at recording new words „in statu nascendi“
• definition of „neologism“: new, but not yet fully lexicalized lexemes

Corpus Linguistic Approach:

• automatic comparison of a current word list with a reference word list built from older corpora and previous word lists
• subsequent lexicographic decision on which words from the generated candidate list are proper neologisms according to the used definition

www.wortwarte.de
3. FINDING NEOLOGISM CANDIDATES

How to find candidates for the „Neologismenwörterbuch“?
1. Editorial evaluation of print and online media
2. Quantitative corpus-linguistic method

Difficulties
• subjectiveness (for 1.)
• finding new meanings (for 2.)
• differentiation between neologisms (in our definition) and ad-hoc formations and nonce words (for 2.)
3. FINDING NEOLOGISM CANDIDATES

Quantitative method for „Neologismenwörterbuch“

• designed to identify the neologisms which are associated with one specific decade, and which are already advanced in their lexicalization process
• comparison of frequency timelines of all words in corpus data (DEREKO – Deutsches Referenzkorpus) from two adjacent time periods A and B
• words in the more recent period B which exhibit a typical timeline are subject to further filtering processes (reduction of remaining non-neologisms such as names and regionalisms)
• result: “neologism candidate list” for inspection by lexicographers
3. FINDING NEOLOGISM CANDIDATES

Recent application of our method:

- aiming at detecting neologisms for the current decade
- corpus spanned 2000-2009 (period A) and 2010-2015 (period B)
- corpus contained over three billion word form tokens yielding around 10 million different word form types
- resulting candidate list (2016) contained 5.483 word form types
4. EVALUATION OF THE QUANTITATIVE METHOD

Linguistic evaluation:

- annotation of the first 500 candidates by lexicographers; categories:
  - proper names
  - semantically transparent
  - inflectional form
  - spelling variant
  etc.

- only the first 500 candidates, because further down the list, the number of proper names and semantically transparent lexemes increases considerably, while the frequency of each candidate in our corpora decreases
4. EVALUATION OF THE QUANTITATIVE METHOD

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Already in “Neologismenwörterbuch”</th>
<th>Category</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pegida</td>
<td>no</td>
<td>proper name</td>
<td>name of political group</td>
</tr>
<tr>
<td>dapd</td>
<td>no</td>
<td>other</td>
<td>abbreviation</td>
</tr>
<tr>
<td>Fracking</td>
<td>yes</td>
<td>-</td>
<td>‘hydraulic fracturing’</td>
</tr>
<tr>
<td>iPad</td>
<td>no</td>
<td>proper name</td>
<td>product name</td>
</tr>
<tr>
<td>Varoufakis</td>
<td>no</td>
<td>proper name</td>
<td>family name</td>
</tr>
<tr>
<td>apn</td>
<td>no</td>
<td>other</td>
<td>abbreviation</td>
</tr>
<tr>
<td>Instagram</td>
<td>no</td>
<td>proper name</td>
<td>app name</td>
</tr>
<tr>
<td>Mietpreisbremse</td>
<td>no</td>
<td>semantically</td>
<td>‘political measure for slowing down the increase of rents’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transparent</td>
<td></td>
</tr>
<tr>
<td>Eurokrise</td>
<td>no</td>
<td>semantically</td>
<td>‘crisis because of the weak Euro’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transparent</td>
<td></td>
</tr>
</tbody>
</table>
Contrasting the findings with the existing lexicographical data (example 1)

• extraction of reference list from “Neologismenwörterbuch” containing all 845 simplex words associated with the entries of the current decade 2011-2016
• mapping on 127 base forms in the database: 81 of the word forms = true positives
• these are associated with 51 different base forms in the “Neologismenwörterbuch”

→ recall of 51/127 = 40% in terms of simplex base forms
  (note: lower precision in view of all 5.483 candidates!)
4. EVALUATION OF THE QUANTITATIVE METHOD

Contrasting the findings with the existing lexicographical data (example 2)

• compilation of a smaller reference list with 390 word forms strictly representing
  – base forms
  – inflectional forms of a headword
  – spelling variants of a headword

• list still contained 130 true positives, i.e., for the reduced reference list the recall remained the same by a minor difference

→ words classified by the lexicographers as morphological variants of head words or sense relations of headwords only, but not deserving headword status themselves had ALSO received a secondary status by our quantitative method
4. EVALUATION OF THE QUANTITATIVE METHOD

Contrasting the findings with the existing lexicographical data: Inspection of false negatives (example 3)

3-D-Drucker, Antänzer, Arabellion, Bestellbutton, BFD, Biodeutscher, Blitzmarathon, Blockupy, Bodycam, Boxspringbett, Brexit, BRICS, Bubble-Tea, Bufdi, Buttonlösung, Cakepop, Chia, Chiasamen, Clickworker, Craftbier, Cross-fit, Crowdfunding, Crowdworker, Crowdworking, Cybergrooming, Darknet, Doodle, Doodleliste, Emoji, Entscheidungslösung, ESM, Facebookparty, Fairteiler, Fakeshop, Faszientraining, Femenaktivistin, Fingerwisch, Fiskalpakt, Fitnessarmband, Flexiquote, Flexirente, Flexitarier, Foodtruck, Fotobombe, fracken, Fracking, Freistoßspray, Frutarier, Fukushima-Effekt, Garagengold, Gettofaust, Glamping, Googlebrille, Grexit, GroKo, Guerillastricken, ...

Key:
grey: entries in „Neologismenwörterbuch“ detected by editorial media evaluation
black: entries in „Neologismenwörterbuch“ which are also part of the candidate list
4. EVALUATION OF THE QUANTITATIVE METHOD

Evaluation of false positives and false negatives:
• 51 true positives and 76 false negatives in a reference list of 127 base forms
• 10 false negatives from this list filtered out by the proper name filter
• majority of false positives because the respective form is not identified as a proper name

→ better setting of the name recognition tool needed
5. OUTLOOK

Inclusion of new lemma types into the „Neologismenwörterbuch“

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>semantically transparent lexemes</td>
<td><em>Mietpreisbremse, Eurokrise</em> ...</td>
</tr>
<tr>
<td>synonyms for already existing entries with a significantly lower frequency</td>
<td><em>Generation Y (headword) – Y-Generation and Yps-Generation (synonyms)</em> ...</td>
</tr>
<tr>
<td>extended usage of lexemes</td>
<td><em>teilen: ‘to share’ – extended usage ‘to share in social media’</em> ...</td>
</tr>
<tr>
<td>phrases from other languages</td>
<td><em>never ever, powered by, all you can eat, sharing is caring</em> ...</td>
</tr>
<tr>
<td>proper names which are the base for new lexemes</td>
<td><em>YouTube, WhatsApp, Twitter, Tinder</em> ...</td>
</tr>
<tr>
<td>terminological lexemes entering the general language</td>
<td><em>Koenzym, Coretraining, SWIFT-Code, Prokrastination</em> ...</td>
</tr>
</tbody>
</table>
REFERENCES


VIELEN DANK