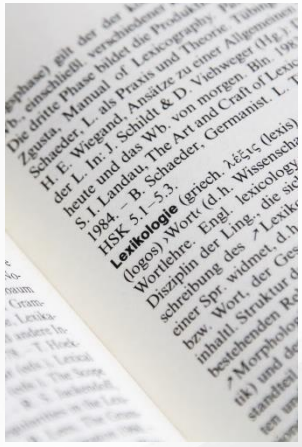
INSTITUT FÜR DEUTSCHE SPRACHE



Annette Klosa-Kückelhaus and Harald Lüngen

NEW GERMAN WORDS: DETECTION AND DESCRIPTION

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

Mitglied der Leibniz-Gemeinschaft

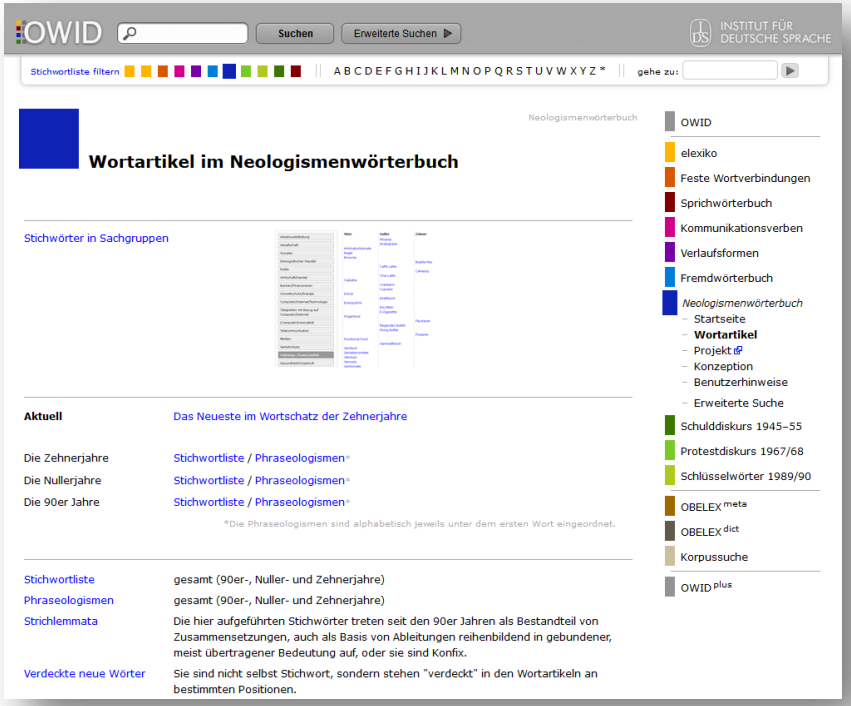
OUTLINE

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

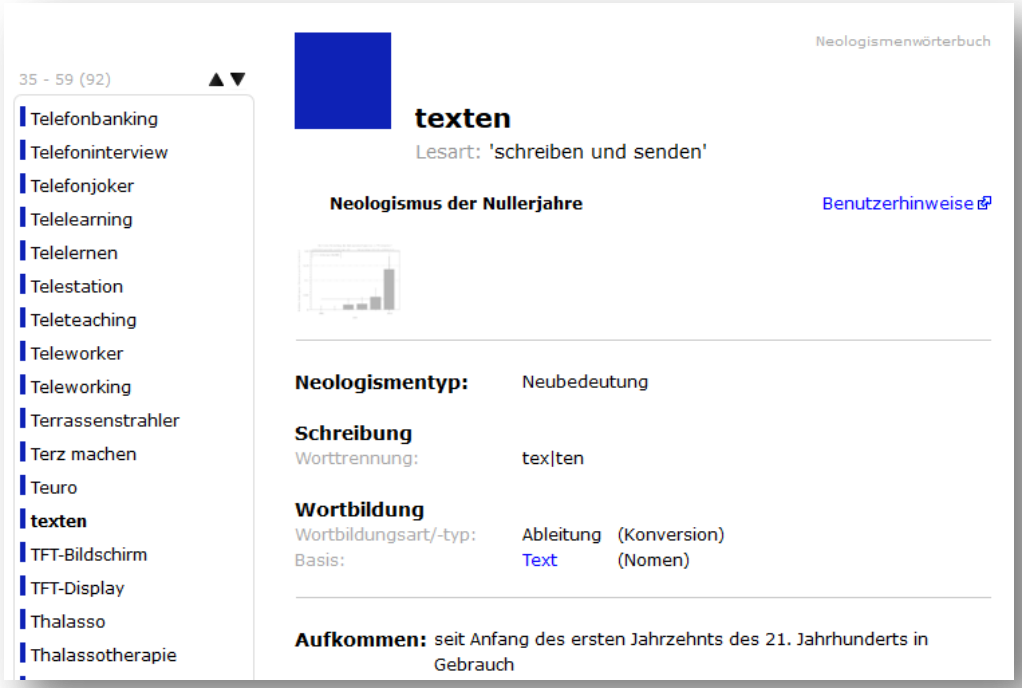


http://www.owid.de/docs/neo/wortartikel.jsp

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"

All entries	**almost 1.900**
Neologisms from 1991-2000	over 1.000
Neologisms from 2001-2010	almost 700
Neologisms since 2011	almost 150
New lexemes	almost 1.550
New elements of word formation	almost 20
New meanings	over 160
New multi-word units	almost 120
Other new lexemes (synonyms, other sense-related words, derivations, compounds, etc.) contained in entries and accessible via list	almost 5.000

1. THE "NEOLOGISMENWÖRTERBUCH"

Lexicographic information

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

Grammatische Angaben

Wortart:	**Verb** (schwach)
Konjugation	
Präteritum:	textete
Partizip Perfekt:	getextet
Perfektbildung:	*mit* haben
	bildbar

Typische Verwendungen

E-Mails texten
der Freundin SMS-Nachrichten texten
die Antwort per SMS texten
mit der Freundin texten
stundenlang texten
per Handy texten
auf Facebook texten

Bedeutungsangabe

einen Text, besonders eine Kurznachricht, in den elektronischen Medien schreiben und versenden

Wortbildungsproduktivität

Präverbfügung: *lostexten, zurücktexten, zutexten*

SMS sind schon fast wieder Steinzeit, mit dem IPhone wird auf Facebook **getextet**, per Twitter werden 140-Zeichen-Messages in den Äther verschickt (Niederösterreichische Nachrichten, 03.02.2010)

http://www.owid.de/artikel/401238?module=neo&pos=13

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.

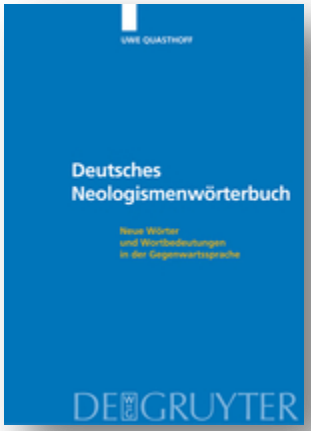
All neologisms in the dictionary are

- fully lexicalized lexemes
- not nonce words.

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

www.wortwarte.de

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

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 defintion) 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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Pegida	14845	0%	0.47625	DE-S	28523828	23955.54	2015	12948	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1897	12948
2	dapd	52487	1.96%	0.54312	DE-S	100781217	21172.48	2010	24053	2012	0	0	0	0	0	0	0	0	0	3	6458	20799	24053	1105	51	18
3	Fracking	7313	55.45%	0.36442	DE-S	14051574	11801.89	2011	2421	2013	0	0	0	0	0	0	0	0	0	0	3	173	1237	2421	2024	1455
4	iPad	10736	6.67%	0.41906	DE-S	20607144	8661.82	2011	2614	2012	0	0	0	0	0	0	1	0	0	0	2160	2362	2614	1410	1181	1008
5	Varoufakis	3890	100.00%	0.46889	DE-S	7474.46	6278.53	2013	3838	2015	0	0	0	0	0	0	0	0	0	0	0	0	48	2	2	3838
6	apn	6776	1.01%	0.54944	DE-S	12999.16	5466.89	2011	6764	2010	0	0	1	0	0	0	0	0	0	0	6764	9	1	0	1	0
7	Instagram	3139	50.98%	0.41574	DE-S	6031.45	5066.71	2012	1531	2015	0	0	0	0	0	0	0	0	0	0	1	2	434	428	743	1531
8	Mietpreisbremse	2995	19.61%	0.5192	DE-S	5754761	4834.35	2014	1483	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	560	952	1483
9	Eurokrise	5219	37.62%	0.39701	DE-S	10007.99	4210.7	2011	1969	2012	1	0	0	0	0	0	0	0	0	0	290	1067	1969	1096	394	402
10	Selfie	2491	6.25%	0.40586	DE-SW	4786348	4021.1	2014	1567	2015	0	0	0	0	0	0	0	0	0	0	0	0	0	78	846	1567
11	Bundesfreiwilligendi	2434	0.99%	0.59285	DE-SW	4676825	3929.12	2011	758	2011	0	0	0	0	0	0	0	0	0	0	68	758	617	342	243	406

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

Candidate	Already in "Neologismenwörterbuch"	Category	Comment
Pegida	no	proper name	name of political group
dapd	no	other	abbreviation
Fracking	yes	-	'hydraulic fracturing'
iPad	no	proper name	product name
Varoufakis	no	proper name	family name
apn	no	other	abbreviation
Instagram	no	proper name	app name
Mietpreisbremse	no	semantically transparent	'political measure for slowing down the increase of rents'
Eurokrise	no	semantically transparent	'crisis because of the weak Euro'

4. EVALUATION OF THE QUANTITATIVE METHOD

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“

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

REFERENCES

Herberg, D., Kinne, M. & Steffens, D. (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. In collaboration with E. Tellenbach and D. al-Wadi. Berlin/New York: de Gruyter.

Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-II* (Release: 01.10.2017). Mannheim: Institut für Deutsche Sprache. Accessed at http://www.ids-mannheim.de/DeReKo [20/03/2018].

Keibel, H., Hennig, S. & Perkuhn, R. (2010). *Effiziente halbautomatische Detektion von Neologismuskandidaten.* Technical Report IDS-KL-2010-01. Mannheim: Institut für Deutsche Sprache.

Neologismenwörterbuch (2006-today), in: OWID – Online Wortschatz-Informationssystem Deutsch. Mannheim: Institut für Deutsche Sprache. Accessed at http://www.owid.de/wb/neo/start.thml [20/03/2018].

Quasthoff, U. (2007). *Deutsches Neologismenwörterbuch. Neue Wörter und Wortbedeutungen in der Gegenwartssprache*. de Gruyter: Berlin.

Wortwarte (2000-today). *Die Wortwarte. Wörter von heute und morgen. Eine Sammlung von Neologismen*. Ed. by Lothar Lemnitzer. Accessed at http://www.wortwarte.de [20/03/2018].

INSTITUT FÜR DEUTSCHE SPRACHE

VIELEN DANK

Mitglied der Leibniz-Gemeinschaft