New German Words: Detection and Description

Annette Klosa, Harald Lüngen
Institut für Deutsche Sprache, Mannheim
E-mail: klosa@ids-mannheim.de, luengen@ids-mannheim.de

Abstract

In this paper, we discuss an efficient method of (semi-automatic) neologism detection for German and its application for the production of a dictionary of neologisms, focusing on the lexicographic process. By monitoring the language via editorial (print and online) media evaluation and interpreting the findings on the basis of lexicographic competence, many, but not all neologisms can be identified which qualify for inclusion in the *Neologismenwörterbuch* (2006-today) at the Institute for the German Language in Mannheim (IDS). In addition, an automated corpus linguistic method offers neologism candidates based on a systematic analysis of large amounts of text to lexicographers. We explain the principles of the corpus linguistic compilation of a list of candidates and show how lexicographers work with the results, combining them with their own findings in order to continuously enlarge this specialized online dictionary of new words in German.

Keywords: detection of neologisms, description of neologisms, corpus linguistics, lexicography

1 Introduction

Entries on new words in general dictionaries of German or in specialized dictionaries of neologisms in German attract many users, as new words pose problems regarding their meaning and usage, their grammatical features, their orthography, and their pronunciation. Neologisms are also often subject of language criticism in German: while some speakers consider them as unnecessary additions to German mainly taken from other languages (with English often as the most common donor language), others realize their importance in filling lexical gaps or providing means of enriching the language with a multitude of possibilities of expression. Generally, language change, which we can see most prominently at the lexical level, is a topic of high interest for all speakers (cf. Lemnitzer 2010: 65).

Therefore, a central issue in lexicography (for German as well as other languages) is to find new lexemes and to identify new meanings for existing lexemes. By monitoring the language via editorial media evaluation and interpreting the findings on the basis of lexicographic competence, many but not all neologisms can be identified. Only automated methods of corpus linguistics can provide a systematic analysis of large amounts of text, offering neologism candidates to lexicographers. In this paper, we discuss our method of neologism detection for German and its application for the production of a dictionary of neologisms in this language.

In the following (cf. Section 2), we present the *Neologismenwörterbuch* (2006-today; cf. Steffens 2017) at the Institute for the German Language in Mannheim (IDS). In Section 3, we discuss related work (other German dictionaries of new words and other methods of detection of neologisms). We describe our semi-automatic method of detection of neologism candidates in Section 4, which we the evaluate in Section 5. In this section, we also focus on the impact of this method for both our dictionary as well as the corpus tool. In a short outlook (cf. Section 6), we then discuss the possibilities of a more extensive use of corpus linguistic findings in our online dictionary of neologisms in the future.

2 The Neologismenwörterbuch

Dictionaries of new words are specialized dictionaries describing the meaning and usage of lexemes in a specific language which became part of the vocabulary at a certain time (for more details cf. Barnhart & Barnhart 1990; Lemnitzer 2010; Wiegand 1990). The *Neologismenwörterbuch* published online by IDS Mannheim is a typical example of this type of dictionary. It covers new words and new meanings established in the past thirty years. The online publication in the dictionary portal OWID (Online-Wortschatz-Informationssystem Deutsch) at IDS Mannheim allows for the continuous addition of new entries. For the decades of 1991-2000 and 2001-2010, print dictionaries are also available (Herberg et al. 2004; Steffens & al-Wadi 2015). The lexicographic concept for these dictionaries goes back to the late 1980s (cf. Heller et al. 1988; Kinne 1989) and 1990s (cf. Herberg 1997 and 1998), when German lexicography had not yet embraced the potential of corpus linguistics. Only in the first decade of the 21st century was a corpus linguistic method of neologism detection developed to supply lexicographers working on the *Neologismenwörterbuch* with candidates for inclusion in the dictionary in combination with candidates taken from the project's own editorial media evaluation (cf. Section 4.1).

The *Neologismenwörterbuch* comprises entries on single words (e.g., *Avatar*), multi-word expressions (e.g., *in der Pipeline*), and new elements of word formation (e.g., *[...]holic*). Not only new words, but also new meanings for existing words in German are described (e.g., *texten* 'send a (short) text message in electronic media'). Proper names are basically excluded from the lemma list in the *Neologismenwörterbuch*; only derivatives with a proper name as their base are included in the dictionary, for example *twittern* ('to send a Twitter message'), but not *Twitter*. As many compounds and derivatives in German are semantically transparent, i.e., can be fully interpreted based on the knowledge of the meaning of their components, they are not lemmatized in the *Neologismenwörterbuch*, either, for example *Eurokrise*, 'crisis because of the weak Euro'.

Lexicographic information comprises etymology, orthography, pronunciation, meaning, usage, grammar, word formation, encyclopedic information, illustrations, and frequency in the corpus. The dictionary aims at covering all neologisms established throughout the last two as well as the current decade, describing each neologism as new for each decade, respectively. Table 1 gives information on the number of entries in the *Neologismenwörterbuch* in March 2018. All entries in the dictionary meet the following definition of "neologism" in our project: 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. To sum up: neologisms in our project are not nonce words, but are defined as fully lexicalized lexemes. Thus, only in retrospect it is possible to decide which words are neologisms and which are not.

Table 1: Numbers of entries in the *Neologismenwörterbuch* in March 2018

All entries	more than 1.800
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

Our definition contains several criteria which cannot be easily operationalized: How do we measure whether a new word is generally accepted or that a majority of speakers perceives it as new? For each neologism candidate, a decision on its possible inclusion in the dictionary has to be based on an individual analysis of the data available. Not only do we look at the number of years and/or months since the lexeme has shown up in the German corpora and the development of its frequency, but also at the way in which it is being used. There are several textual indicators for words which are not yet fully lexicalized (cf. Lemnitzer 2010: 69): quite often, they are used in quotation marks or are followed by short definitions. In particular words borrowed from other languages initially do not exhibit a full declination paradigm in German; nouns often show different genders, before they settle for one grammatical gender. Pronunciation as well as orthography show a lot of variation in the beginning as well. Moreover, only fully lexicalized words in German can enter into word formation products in combination with Germanic as well as loan morphemes. Candidates for inclusion in the dictionary, whether from our editorial reading or detected automatically, are evaluated according to these criteria.

3 Related Work

3.1 German Dictionaries of New Words

German dictionaries of new words are a fairly recent development: the first German print dictionary of neologisms was published at the beginning of the new century at IDS Mannheim (Herberg et al. 2004), shortly followed by the *Neologismenwörterbuch* online (in 2006; cf. section 2). In 2007, a first, strictly corpus-driven print dictionary (Quasthoff 2007) was published. It contains almost 2,300 entries, giving short definitions, corpus examples, information on the subject area, and frequency diagrams for each head word. Neologisms in this dictionary are words whose frequency has increased significantly between 2000 and 2006. However, not all of these are proper new words according to our definition, as quite a number of the entries comprise nonce words

Since 2000, the website Wortwarte (2000–today) has collected and published German neologisms automatically extracted from newspaper web pages and other online sources. Here, neologisms are defined as new, but not yet fully lexicalized lexemes. The Wortwarte-dictionary aims at recording new words "*in statu nascendi*" (Lemnitzer 2010: 67). In 2017, for example, approximately 2,900 new words were registered with short grammatical information and one corpus example.

Of course, general dictionaries of German also update their lists of headwords, adding new entries for new editions. The latest edition of *Duden – Die deutsche Rechtschreibung* (2017) contains more than 5,000 new entries (as part of 145,000 entries in total) according to Duden publishing house. Many of these are not neologisms in our definition, but, for example, transparent compounds or proper names.

3.2 (Semi-)Automatic Detection of New Words

Most approaches to a corpus-based detection of neologisms use press corpora and to some extent also web-specific corpora (Lemnitzer 2010; Quasthoff 2007). Wortwarte (2000–today) automatically compares a current word list (i.e., a word list of the day derived from press and web texts) with a reference word list built from older corpora and previous word lists, thus obtaining a list of new words of the day. Subsequently, a lexicographer decides which words from the generated candidate list is a proper neologism according to the used definition. Following Falk et al. (2014), this approach may be classified as based on exclusion lists.

Another type of approach uses corpora which are partitioned into sub-corpora by year, and then tries to identify typical frequency timelines of neologisms, with a rise of frequency in the present and with minimum frequency conditions to model its establishment in the language and to exclude nonce words. Such an approach has been used by Quasthoff (2007), and our approach falls into this category, too, cf. Section 4.2. An alternative approach is exemplified by Falk et al. (2014), who combine exclusion lists with a supervised machine-learning approach in which typical features of the linguistic context of neologisms are extracted from (French) press corpora. The authors note that the advantage of their approach is that it does not heavily rely on large, diachronic, time-annotated corpora.

All of these approaches are semi-automatic in that candidate lists are generated which must be post-processed manually in order to select the proper neologisms according to the respective definitions.

4 Finding Neologism Candidates

4.1 Editorial Evaluation of Print and Online Media

Since starting the *Neologismenwörterbuch*, lexicographers working on the dictionary have collected candidates for new words through intensive daily browsing through a number of print and (later also) online media. Candidates are entered into a database with information on the date of the first sighting (or hearing), the source, and a first frequency result of a search in our corpus or on the web. A tentative short definition and links to other sources are added, if possible. All candidates are classified according to the decade they entered the German language. Some candidates remain in the database for further monitoring for several years, other candidates directly become entries in the dictionary.

In addition to the dictionary staff, a small number of users of the *Neologismenwörterbuch* send in suggestions for new entries. In the future, we plan to systematize collaboration with users in the editorial evaluation of print and online media by introducing an appeal to readers to send in their findings via a form.

4.2 Quantitative Method

Our method for the quantitative detection of neologism candidates was originally developed by Holger Keibel and described in detail in the technical report Keibel et al. (2010). According to the setting of the *Neologismenwörterbuch*, it is designed to identify neologisms which are associated with one specific decade, and which are already advanced in their lexicalization process, i.e., excluding ad-hoc formations and nonce words. To achieve this, frequency timelines of all words are compared in corpus data from two adjacent time periods A and B. Those words in the more recent period B which exhibit a typical timeline are subject to further filtering processes aimed at reducing the remaining non-neologisms such as names and regionalisms; the resulting list is considered the “neologism candidate list” that should be more closely inspected by the lexicographer.

In the recent application of the method that we evaluated for this paper, we have compared the corpus data from the period 2010-2015, representing the current decade, with data from the previous decade 2000-2009 in order to discover neologism candidates of the current decade. (The delimitation of decades in this application is the one originally applied by Keibel et al. (2011) and deviates slightly from the one used in the *Neologismenwörterbuch* (2001-2010 and 2011-2020). This was adjusted in later applications.)

4.2.1 Corpus and Frequency List

As corpus data, we use a virtual corpus with over three billion tokens of press text which in the application described in Section 4.2.3 and evaluated in Section 5, spanning the years 2000-2015, and which is derived from the German Reference Corpus DeReKo. DeReKo is hosted by IDS and is the largest linguistic text archive for the German language, currently with 42 billion tokens (Institut für Deutsche Sprache 2017). The bulk of DeReKo is made up of press corpora, but many other genres such as fiction, science, specialized texts, debate protocols, and computer-mediated communication, are represented, too. The newspaper titles represented in the current project corpus were selected because: a.) most of them are available in DeReKo for most of the years in the period under scrutiny, and b.) they are well distributed across the four language regions North, East, South, and Southwest of Germany.

Initially, a huge frequency list of all word forms occurring in the project corpus representing both periods A and B is built. The quantitative method strictly investigates word forms only, i.e., unclassified corpus tokens. Base forms or lemmata are not considered, primarily because automatic lemmatization tools will frequently fail to lemmatize new and unknown word forms, but also because it is interesting to keep track of possibly differing frequency properties of the different inflectional forms or spelling variants.

4.2.2 Filtering of the Frequency List

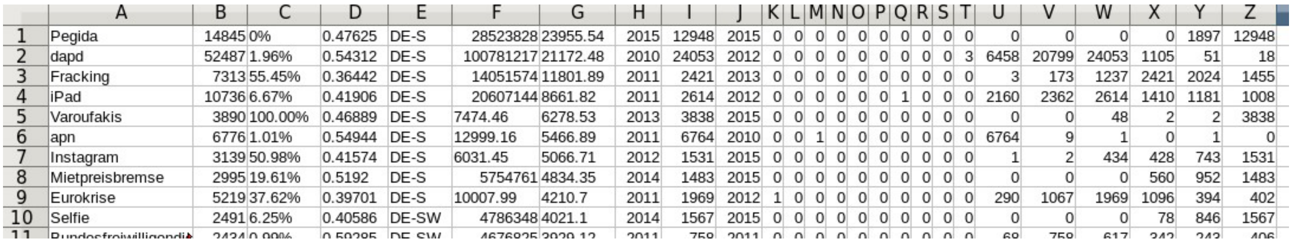
From the full frequency list, obvious non-lexical items such as numbers or URLs are identified based on graphematic criteria and removed in a first filtering step. The second step consists of filtering out all words with an overall absolute frequency below a certain minimum (currently 10). In a third step, those forms are filtered out that do not conform to a set of quantitative criteria which, amongst other things, define a maximum frequency in period A, a minimum frequency in period B, and minimum frequencies for the years after the year of appearance and the peak year. These criteria are to characterize the typical timeline of a neologism of period B and to filter out other words, including short-lived, ad-hoc, non-lexicalized formations and usages.

The resulting list still contains many names and regionalisms that conform to the quantitative criteria; hence in the fourth filtering step we try to filter out regionalisms by removing word forms that show a bias of occurrence for one of the four sub-corpora representing German language regions according to the DP (deviation of proportions) dispersion measure by Gries (2008). In a fifth filtering step, names are filtered by sending KWIC result lines with candidates derived from the corpus to the Stanford Named Entity Recognizer (Finkel et al. 2005). Filtering is applied in a conservative fashion, because in the lexicographic context a maximization of recall is considered more important than a maximization of precision. The latter would bear a higher risk of losing relevant candidates in the filtering processes.

The resulting list is our “neologism candidate list”, which still contains many obvious non-neologisms, mostly names that had not been correctly identified by the NER tool. Keibel et al. (2010) point out that these could be filtered efficiently manually even by someone who is not an expert before the final candidate list is analyzed by a lexicographer.

4.2.3 Recent Application and Candidate List

In one of our recent applications of the method to detect neologisms of the current decade, the corpus spanned 2000-2009 (period A) and 2010-2015 (period B) and together contained over three billion word form tokens yielding around 10 million different word form types. The resulting CANDIDATE-LIST_2016 contained 5,483 word form types (neologism candidates).



	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

Figure 1: Top of CANDIDATE-LIST_2016, including frequencies by year, LLF, M-score (cf. Keibel et al. 2010), peak year, probability of being a name or a regionalism, and ranked according to the M-score.

The initial word lists derived from our corpora and the full resulting CANDIDATE_LIST_2016 will shortly be published on the research data server of our institution[1] so that others can try out alternative approaches on the same data.

5 Evaluation of the Quantitative Method

5.1 Linguistic Evaluation of the Candidate List and Comparison with Editorial Media Evaluation

Lexicographers working on the *Neologismenwörterbuch* annotated the first 500 false positives (FPs) in CANDIDATE-LIST-2016 with the six categories *Proper Name, Semantically Transparent, Occult Neologism, Inflectional Form, Spelling Variant,* and *Other*. Table 2 shows an extract of annotated datasets 1-15 from the CANDIDATE-LIST_2016 with explanatory comments.

Among the top 500, 219 of the false positives were annotated as proper names, 179 as semantically transparent, 79 as "other", 13 as inflectional forms, and 10 as occult neologisms. CANDIDATE LIST_2016 was annotated while large parts of the entries for the first decade and the first half of the second decade of the 21st century had already been published online. Thus, quite a number of candidates were already contained in the *Neologismenwörterbuch* either as a full lemma or as a sub-lemma (e.g., as a word formation product, or as a less frequent synonym).

The majority of candidates which had not already been found through editorial media evaluation and accordingly included in the dictionary are either proper names (e.g., *Instagram*) or semantically transparent lexemes (e.g., *Eurokrise*). At the moment, proper names are basically excluded from the lemma list in the *Neologismenwörterbuch*; only derivatives with a proper name as their base are included in the dictionary, for example *twittern* ('to send a Twitter message'), but not *Twitter*, *Youtuber* ('somebody watching or somebody producing YouTube videos'), but not *Youtube*. As many compounds and derivatives in German are semantically transparent, i.e., can be fully interpreted based on the knowledge of the meaning of their components, they are not lemmatized in the *Neologismenwörterbuch*, either, for example *Eurokrise* 'crises because of the weak Euro'. Only the top 500 words were annotated in CANDIDATE-LIST_2016, 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.

[1] https://repos.ids-mannheim.de/

Table 2: Extract from annotated list of candidates for dictionary of neologisms; abbreviations: pn = proper name, st = semantically transparent; oth = other (e.g., abbreviations)

Candidate	**Lemma/Sub-lemma in Dictionary**	**Category**	**Comment**
Pegida	no	pn	name of a political group
dapd	no	oth	abbreviation
Fracking	yes		'hydraulic fracturing'
iPad	no	pn	product name
Varoufakis	no	pn	family name
apn	no	oth	abbreviation
Instagram	no	pn	app name
Mietpreisbremse	no	st	'political measure for slowing down the increase of rents'
Eurokrise	no	st	'crisis because of the weak Euro'
Selfie	yes		engl.: *selfie*
Bundesfreiwilligendienst	no	pn	name of German nationwide voluntary service
Grexit	yes		engl.: *grexit*
iOS	no	oth	abbreviation
Fiskalpakt	yes		'fiscal pact'
Kobane	no	pn	geographical name

Therefore, these candidates are presumably less relevant for inclusion in our dictionary. As of 2017, a candidate list has ben compiled and evaluated annually so that the number of words which are already part of the dictionary should decrease in the future. The editorial media evaluation is still needed to supplement this method, because we also include multi-word expressions, new elements of word formation, and new meanings in the dictionary which currently cannot be found automatically. Also, many new lexemes in our dictionary had not been discovered by the corpus-linguistic method explained above (false negatives). The following list gives an overview of entries from the first half of the second decade of the 21st century in the *Neologismenwörterbuch* showing which ones were detected in our editorial media evaluation (grey) and which ones were also contained among the top 500 of the automatically compiled CANDIDATE-LIST_2016 (black):

- New lexemes: *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, Hashtag, Helikoptereltern, Hipsterbart, Homestaging, Hugo, Hygieneampel, Inklusionsklasse, Jahnbehörde, Kampfradler, Keniakoalition, Kinesiotape, Kryptohandy, Kryptoparty, leaken, Leo, Like, liken, Loopschal, Memoriamgarten, Mikrojob, Mingle, Netzpartei, Occupybewegung, Pflege-Bahr, Phablet, Pinkifizierung, Pop-up, Pop-up-Restaurant, QR-Code, Reichweitenangst, Repaircafé, Retweet, retweeten, Scamming, Selfie, Selfiestick, Seniorazubi, Sexting, Shapewear, Shitstorm, Smart-TV, Smartwatch, Spotted-Seite Stadtgärtnern, Streetfood, Strickgraffito, Strickguerilla, Superfood, tinder, Tofutier, Upcycling, Vatileaks, veggie, Veggieday, Veggietag, Vöner, Webinar, WhatsApp ('short message'), whatsappen, Willkommensklasse*

- New meanings: *aufpoppen, Computeruhr, dampfen, Dampfer, Energiearmut, Flugmodus, Loop, stromern, Tunnel, wischen*
- New multi-word expressions: *"Gefällt mir", "Gefällt mir"-Button, arabischer Frühling, falsche Neun, falscher Neuner, grüner Smoothie, "hätte, hätte, Fahrradkette", Natural Running, Second Screen, vertrauliche Geburt, ziemlich beste [X]*

These findings by our lexicographic team shall lead to finding better parameter settings to use with the quantitative method for the detection of neologism candidates. The majority of the false positives clearly arise because the respective form is not identified as a proper name by the NER tool. We shall try to find a better setting of the tool in new experiments, but at the same time we expect this problem to be hard to eliminate. Note that semantically transparent words would count as proper neologisms in other approaches to automatic detection (cf. Section 3.2). Note also that inflectional forms of head words should rather count as true positives from the perspective of the quantitative method.

5.2 Evaluation of the Filter Criteria

We also extracted a reference list from the underlying database of the *Neologismenwörterbuch* containing all simplex words associated with the entries of the current decade 2011–2016, which as explained above had originally been compiled solely through the editorial media evaluation. This reference list contained 845 word forms which were mapped on 127 base forms in the database. Eighty-one of the word forms were true positives, i.e., also contained in our CANDIDATE-LIST_2016, and these in turn were associated with 51 different base forms in the database. Thus our CANDIDATE-LIST_2016 yielded a recall of 51/127 = 40% in terms of simplex base forms. In view of the number of 5,483 items on our Result List 2016, the precision is of course far lower.

We removed those items from the reference list that, according to the database, represented sense relations of head words (e.g., *Radrowdy* 'bike rowdy' as a synonym of the headword *Kampfradler* 'bike rowdy'), and those that represented word formations derived from or composed with head words (e.g., *Kampfradlerin* as a derivative of *Kampfradler*), and obtained a smaller reference list with 390 word forms strictly representing only either the base form, an inflectional form, or a spelling variant of a headword. This reduced reference list still contained 130 true positives, i.e., for the reduced reference list the recall remained the same by a minor difference. From this calculation we concluded that those words that had been classified by the lexicographers as morphological variants of head words or sense relations of headwords only, but not deserving headword status themselves, had received a secondary status by our quantitative method, too, which was a nice confirmation.

Next we wanted to know whether we would still obtain a reasonable recall if we cut off our ranked candidate list at some point. It turned out that if we cut of the list at 4,000 entries it would still contain 78 of the 81 true positives. Cutting of the CANDIDATE-LIST_2016 at a higher point would deteriorate the recall, e.g., 69 true positives remaining when cutting of at rank 2500. We thus concluded that ⅘ or more of the CANDIDATE-LIST_2016 would have to be taken into consideration in order to find nearly all neologisms contained in it. These figures were confirmed in later applications.

Another form of evaluation was a closer inspection of the False Negatives (FNs) to find out when and why they were falsely filtered out, with the ultimate goal of improving the performance of the filters. As pointed out above, FNs (i.e., type II errors) imply a lower recall and are considered more severe errors in the lexicographic context than false positives. Having identified 51 true positives in a reference list of 127 base forms, the number of false negatives (FN) was 76. It turned out that eight of the FNs had not occurred in the corpus in the first place, one had been filtered out by the graphemic cleaning procedure, eleven had been filtered because their absolute corpus frequency was under the required minimum of ten corpus hits, 44 were filtered out by the complex quantitative criterion, two were filtered out by the regional dispersion filter, and finally, ten were removed by the proper name filter, cf. Table 3 for details.

The distribution illustrated in Table 3 shows that most (44) of the FNS are falsely filtered out by the complex quantitative criterion. Step 0 and Step 2 falsely filtered out 19 FPs – for these, we assume that the editorial media evaluation program had been ahead of the corpus-based method, i.e., the lexicographers had already discovered them before they showed up in our press corpora at a reasonable number of occurrences. We would assume that they will be included in a future candidate list when the corpora are extended by data from beyond 2015. Ten FNs were filtered out because they were falsely analyzed to be proper names by the named entity recognizer.

Table 3: When were the false negatives, i.e., neologisms that were in the corpus, but did not make it in the final CANDIDATE-LIST_2016, filtered out?

Filtering Step		**#**	**FNs removed (as base forms)**
Step 0	Not in corpus	8	*3-D-Drucker, Doodleliste, Fukushima-Effekt, Jahnbehörde, Pflege-Bahr, Pop-up-Restaurant, QR-Code, Seniorazubi*
Step 1	Graphemic cleaning	1	*ESM*
Step 2	Absolute frequency < 10	11	*Chiasamen, Fakeshop, Garagengold, Gettofaust, Gruselclown, Guerillastricken, Mikrojob, Selfiestick, Strickgraffito, tindern, whatsappen*
Step 3	Quantitative filter	44	*aufpoppen, Bestellbutton, Biodeutscher, Bubble-Tea, Buttonlösung, Chia, Clickworker, Craftbier, Cybergrooming, dampfen, Dampfer, Doodle, Energiearmut, Facebookparty, Flexirente, Fotobombe, Googlebrille, Hipsterbart, Hugo, Keniakoalition, Kinesiotape, Kryptohandy, Leo, Like, Loop, Mingle, Netzpartei, Pinkifizierung, Pop-up, Reichweitenangst, Sexting, Shapewear, Spotted-Seite, Stadtgärtnern, Streetfood, Strickguerilla, stromern, Tofutier, Tunnel, Upcycling, veggie, Vöner, Webinar*
Step 4	Regional dispersion	2	*Antänzer, Memoriamgarten*
Step 5	Proper name	10	*Arabellion, BFD, Blockupy, BRICS, Bufdi, Darknet, Occupybewegung, Vatileaks, Veggieday, Whatsapp*
		Σ = 76	

6 Conclusion and Outlook

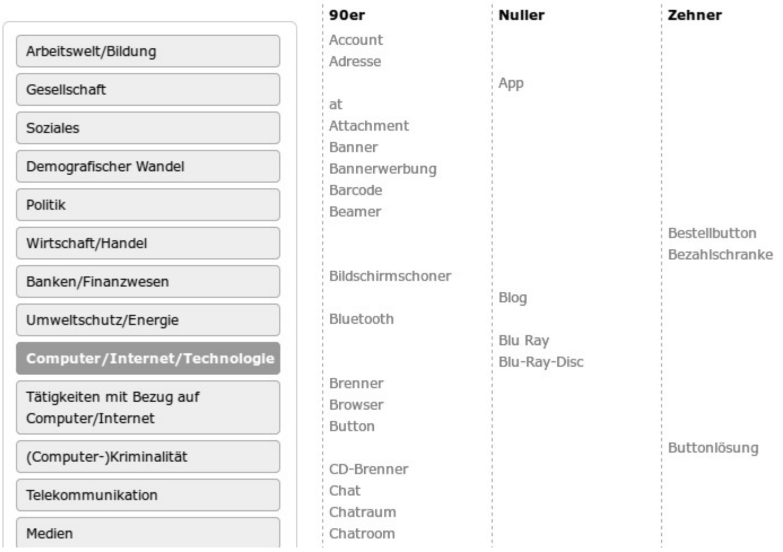
With respect to our quantitative method for detecting new words, the majority of false positives clearly occur because the respective form is not identified as a proper name by the NER tool. On the other hand, several false negatives were not found because they are falsely filtered out as proper names by the NER tool. We shall try to find a better setting of the NER tool in further experiments, but at the same time we expect this problem to be hard to eliminate using current NER technology.

The majority of false negatives are filtered out by the complex quantitative criterion describing the typical timeline of a lexicalized new word (cf. Table 3). This is contrasted by the already fairly low precision (more than 5,000 candidates) our application yielded. Nevertheless, we shall try to improve the quantitative filter settings through more experiments and by carefully adjusting them for each new application, i.e., new corpora and corpus partitions.

As shown in Table 2, many neologism candidates are semantically transparent compounds or derivatives. Instead of excluding these completely from our dictionary, we consider describing them in short entries in the future with the following (reduced) lexicographic information: orthography, pronunciation, word formation, grammar, definition, a small number of corpus examples, encyclopedic information and illustration where necessary or possible. Short entries are also considered for other candidates that could enhance the *Neologismenwörterbuch* in the future:

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

Many of these possible new short entries for the *Neologismenwörterbuch* belong to the following discourses: sports, society, political measures, economy, media, transport, crime, and terrorism. As the *Neologismenwörterbuch* online already offers access to all entries via subject areas (cf. Figure 2), the expansion of the dictionary in this direction seems worth paying attention to.



90er	Nuller	Zehner
Account		
Adresse		
	App	
at		
Attachment		
Banner		
Bannerwerbung		
Barcode		
Beamer		
		Bestellbutton
		Bezahlschranke
Bildschirmschoner		
	Blog	
Bluetooth		
	Blu Ray	
	Blu-Ray-Disc	
Brenner		
Browser		
Button		
		Buttonlösung
CD-Brenner		
Chat		
Chatraum		
Chatroom		

Figure 2: Access to entries in the *Neologismenwörterbuch* via subject area and listing of entries in their decade of emergence (http://www.owid.de/docs/neo/gruppen.jsp)

References

Barnhart, R., Barnhart, C. (1990). The Dictionary of Neologisms. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (eds.) Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter, pp. 1159-1166.

Duden – Die deutsche Rechtschreibung (2017). 27th edition. Ed. Dudenredaktion. Berlin: Dudenverlag.

Falk, I., Dernhard, D. & Gérard, C. (2014). From Non Word to New Word: Automatically identifying Neologisms in French Newspapers. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC, The 9th edition of the Language Resources and Evaluation Conference*, *May 2014, Reykjavik, Iceland.*

Finkel, J. R., Grenager, T. & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, USA, 2005.*

Gries, St. Th. (2008). Dispersions and adjusted frequencies in corpora. In *International Journal of Corpus Linguistic*s, 13(4), pp. 403-437.

Heller, K., Herberg, D., Lange, C., Schnerrer, R. & Steffens, D. (1988). *Theoretische und praktische Probleme der Neologismenlexikographie. Überlegungen und Materialien zu einem Wörterbuch der in der Allgemeinsprache der DDR gebräuchlichen Neologismen.* Berlin: Zentralinstitut für Sprachwissenschaft.

Herberg, D. (1997). Neologismen im allgemeinen Wörterbuch oder Neologismenwörterbuch? Zur Lexikographie von Neologismen. In K.-P. Konerding, A. Lehr (eds.) Linguistische Theorie und lexikographische Praxis. Symposiumsvorträge. Heidelberg 1996. Tübingen: Niemeyer, pp. 61-68.

Herberg, D. (1998). Auf dem Weg zum deutschen Neologismenwörterbuch. In A. Zettersten, V. H. Pedersen & J. E. Mogensen (eds.) Symposium on Lexicography VIII. Proceedings of the Eighth International Symposium on Lexicography May 2-4, 1996, at the University of Copenhagen. Tübingen: Niemeyer, pp. 187-192.

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.

Kinne, M. (1989). Endlich: Ein deutsches Neologismenwörterbuch. In *Der Sprachdienst*, 4, pp.115-116.

Lemnitzer, L. (2010). Neologismenlexikographie und das Internet. In: *Lexicographica*, 26, pp. 65-78.

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.

Steffens, D. (2017): Vom Print- zum Onlinewörterbuch – Zur Erfassung, Beschreibung und Präsentation von Neologismen am IDS. In J. Dąbrowska-Burkhardt, L. M. Eichinger & U. Itakura (eds.) Deutsch: lokal – regional – global. Tübingen: Narr, pp. 281-294.

Steffens, D., al-Wadi, D. (2015). *Neuer Wortschatz. Neologismen im Deutschen 2001-2010*. Mannheim: Institut für Deutsche Sprache.

Wiegand, H. E. (1990). Neologismenwörterbücher. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (eds.) Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter, pp. 2185-2187.

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].