Making sense of nonce words

By *Lothar Lemnitzer*

Abstract: The project „Wortwarte“ (www.wortwarte.de) aims at the collection and documentation of new words in German. The collection is based on German newspaper texts and enables lexicologists and lexicographers to describe lexical aspects of language change. One of the shortcomings of the approach is that only words can be detected on the basis of their form. New meanings of existing words are this invisible. In this paper it will be argued that a retrospective analysis of the collected data, in particular of the huge number of German compounds, allows insights into meaning changes of the parts of these compounds, in particular of the morphological head of such constructions. I will illustrate this point with three examples. The paper ends with some perspectives on the potential to automate parts of these analyses.

Introduction

The data on which this article is based originate in the project „Die Wortwarte“. The outcome of the project is a web-based dictionary of new German words which is updated on a daily basis. Currently (i.e. July 2011) nearly 40 000 words are registered. These words have been collected since September 2000. For every word, there is a rudimentary article with some grammatical information and one usage example, but no definition. Links to Google and Wikipedia make it easy for the reader to collect further usage examples or even to consult a full-fledged encyclopedia entry. The term “Wortwarte” has been chosen to allude to a meteorological observatory. Such a device samples and collects meteorological data on a regular basis. It does not have to say anything about climate change. Nevertheless, the data which are collected at many observatories and over a long period allow scientists to test their models and hypotheses about the development of the climate. Similarly, the “Wortwarte” provides data about lexical innovations. It does not say anything about trends in lexical language change, but the collected data will hopefully enable linguists lexicologists to test their hypotheses about this aspect of language change, cf. Lemnitzer (2007)

The following features of the “Wortwarte” are relevant for the rest of the paper:

- The „Wortwarte“ presents information about ca. 40 000 new words which have been collected over the last eleven years; In the following, I will call this list of headwords “list1”.

- This collection is based on a monitor corpus (i.e. a corpus which is constantly updated to track rapid lexical change; cf. Sinclair (1982)) which is in its majority fed by the online texts of some daily and weekly German newspapers (e.g. Süddeutsche Zeitung and Die Zeit). These texts are being downloaded daily. After segmentation, linguistic analysis of the individual words and the preparation of concordance lines for new words the full texts are removed from the local disc, mainly to avoid a breach of Intellectual Property Regulations.

- The 40 000 headwords are a manual selection from the 6 million „new” strings which appear to be “new” in the sense that they have not been recorded yet, neither in a German reference corpus nor in the word lists which have been collected so far. This constantly growing word type list will be called “list2” in the following. Not all of these strings are “words” in the proper sense of this term; many new strings are simply results of typing errors or errors in rendering the text on the web site.

- Still the major part of the words, in the proper sense of this term, is so-called nonce-words. They are created by an author for one single communicative task. Most of these words are used never again. The German language with its tendency for long and complicated compounds (e.g. *Telekommunikationsvorratsdatenspeicherung, Materialkombinationsparameterzulässigkeitsgrad*) lends itself to the spontaneous creation of nonce words. Since the Wortwarte observes and collects such words in statu nascendi, it cannot be determined at that point in time whether a word will be dropped and remain a nonce word or be lexicalized and become a part of the lexicon of the German language.

- One of the technical limits of the „Wortwarte“ is it inability to detect new senses of existing words. The collection process is purely form-based and lacks any “semantic intelligence”. This reflects the state of the art of natural language processing, and for at least the next few years this limitation has to be accepted.

To summarize the relevant features of the approach:

- The Wortwarte is close to the language use of the day. New words which have been spread yesterday are analyzed today and will be presented on the website the following day.

- The recording of new words in statu nascendi renders it impossible to forecast the fate of this words: will it become part of the German vocabulary or not. This is in sharp contrast to projects of retrospective neologism lexicography (e.g. Herberg et al. (2004)) where “new” words are described which are already established parts of the vocabulary of the language at the time of their selection and description.

- Finding new meanings of existing words is out of scope of this approach.

However, I will argue in the following that on the basis of the data which have been collected so far, including all the nonce-words, new meanings of existing words can be detected retrospectively.

Detecting new meanings of existings words by means of compound analysis

The following discussion is based on a concept of meaning which is originates in the theoretical framework of the British contextualism. One of its major claims is that, roughly speaking, the contexts in which a word appears, or, more precisely, the accompanying words which form this context, determine the meaning of a word (for a linguistic discussion of the British contextualism, cf. de Beaugrande (1991), ch. 8, for a formal account of such an approach cf. Widdows (2004)). Following this idea, a word which appears in separate classes of contexts has different meanings (or readings). Take the word 'mouse' as an example. If it appears in a context which is characterized by words like 'laboratory', 'study' and 'treatment', it should have another interpretation than the same word appearing in contexts which are characterized by words like 'keyboard' and 'click'. It is another yet interesting issue whether two meanings of a word are related, e.g. by semantic processes like meaning extension by metaphorical use etc.

As has been mentioned before, in German pars of the context of a word can be created by words which are connected with the target word via the morphological process of compounding (e.g. *Labormaus* and *Computermaus*, to extend the example above. This is advantageous for our approach: we are able to collect all compounds of which our target word is a part, analyze them and group the co-occurring compound parts into context classes. These context classes are supposed to determine the meanings or readings of the target word.

These assumptions will be tested in the following with three examples of target words of which we assume that they have assumed additional meanings or readings in the last decades. The examples are *Blase, Heuschrecke* and *Szene.* Translation equivalents for the various readings of these words will be given below.

These three lexical units share the feature that a change in their meaning has occurred in the last decades. These new meanings will be described and it will be shown how these new meanings are related to the established meanings.

First of all, a point of reference is needed, in other words a lexical resource which more or less exhaustively lists all the meanings of a lexical item which are in use a the time of writing of the dictionary article. For several reasons, the „Wörterbuch der deutschen Gegenwartssprache“ has been selected. This six-volume general language dictionary appeared between 1962 (vol. 1) and 1977 (vol. 6). The dictionary comprises around 120 000 entries and was renowned as a ground-breaking lexicographic enterprise at its time. We will refer to this dictionary as WDG in the following. The fact that the six volumes appeared between 1962 and 1977 implies that the German vocabulary is recorded as it was used 40 to 50 years ago. The dictionary is therefore ideal as a blueprint against which the up-to-date us of a word can be documented, on the basis of the aforementioned Wortwarte-data. (the making of the WDG and its impact on the lexicography of contemporary German is described in Kramer (2011)).

The meanings which are registered for the three chosen lexical items in the WDG will be compared to the set of compounds in which the lexical item appears as one part. It will thus be shown that the change in meaning (an extension of the meaning of the word in most cases) can be detected through the analysis of these compounds. I will also sketch the kind of meaning change and the underlying semantic processes where this is possible.

Example 1: „Blase“

The relevant part of the WDG article will be reproduced in the following. The full article can be visited at www.dwds.de, or, to view the authentic WDG style, at retro.dwds.de. The article will be reduced to the definition which is given for each reading. In addition, a translation equivalent for each reading will be given to make the argument easier to follow for the reader with little knowledge of the German language.

Blase

1. mit Luft oder Flüssigkeit gefüllter Hohlraum ['bubble']

a) kleine, halbrunde Wölbung auf der Oberhaut, in der sich Flüssigkeit ansammelt ['blister']

b) kleiner, kugelförmiger, mit Luft gefüllter Hohlraum, Luftblase ['bubble']

c) häutiges Hohlorgan der Menschen, der Tiere, in dem sich Urin ansammelt, Harnblase ['bladder']

2) Gesellschaft, Bande ['gang']

Each reading of the word appears in at least some of the compounds of the Wortwarte data, e.g.:

Reading 1 a) *Grundwasserblase* (= related to liquids which are enclosed)

Reading 1 b) *Hitzeblase* (‘spot on the skin caused by heat’)

Reading 1 c) *Kuhblase, Urinblase, Reizblase* (related to the animal or human bladder)

Reading 2) *Künstlerblase, Politikerblase* (related to groups of people)

Additional readings of this word can be gleaned from other compounds, where the target word is not compatible with one of the readings found in the WDG. Most of these compounds can be assigned to the subject field of economics. In particular, the words with which the target word appears are from the domain of economics and have the following ontological classes:

Dotcomblase (dotcom = institution),

Spekulationsblase, Konsumblase (speculation etc = activity),

Derivatblase (derivative = (financial) product),

Roughly, the (new) economic meaning of the word „Blase“ appears in 70 % of the compounds with *-blase* as head element. This is an indication that a) this meaning of *Blase* is lexicalized and b) that at least in compounds this meaning is the predominant one.

On the basis of these data one could infer this new meaning of *Blase* as „an economic activity performed by economic institutions in which products are involved which becomes too big and consequently gets out of control of the actors, resulting in a burst after which nothing remains or, in other word, everything vanishes in thin air”. This analysis can at least form the basis of a definition of this new reading of the word.

Concerning the process of meaning change, we can infer a metaphorical use of the word *Blase* in its generic sense (sense 1 in the WDG). The relating features which are the basis of the metaphor are the size and fragility of the object and its tendency to burst if it becomes too large or is otherwise damaged. These features are transferred to abstract economic states that are the results of some activities. Complex economic behavior is expressed in a strong image and thus visualized as an object / process which is known to almost everybody.

Example 2: “Heuschrecke”

The change of meaning of the word *Heuschrecke* ('locust') is one of the rare cases where the emergence of the new meaning can be tracked down to a single concrete utterance. This meaning has in this way been established by the German Social Democratic politician Franz Müntefering in an interview with the German newspaper „Bild“ dating back to the 17. April 2005. In this interview, Müntefering addressed private equity companies as „Heuschrecken“.

Here is the entry of the WDG:

Heuschrecke

„in zahlreichen Arten vorkommendes, Wiesen und Laubbäume bewohnendes Insekt, dessen Hinterbeine verdickt und dadurch sehr sprungkräftig sind“ [‘grasshopper‘,‘locust‘]

This original meaning is activated in some of the compounds recorded by the Wortwarte, e.g. *Laubheuschrecke* (=a kind of locust) and *Heuschreckenzüchter* (=a breeder of locusts).

The new meaning which has been coined by Müntefering occurs in the majority of the Wortwarte compounds where the compound refers to the field of activity of the referred actors (*Rohstoffheuschrecke*, dealing with commodities) or to the way in which the related economic activity is organized (*Heuschreckenfonds*, a fund). The share of these examples is around 50 %. In another 40 % of the compounds the discussion which had been ignited by the Müntefering interview is referred to (*Heuschreckendiskussion*, *Heuschreckenkritik*, *Heuschreckenkampagne* (discourse, critique, campaign), etc.). In just less than 10 % of the compounds the original reading is activated.

While the reference to the „Heuschreckendisurs“ (discourse about locust-like behaviour of some capitalists) will disappear in time, the reference to finance capitalists and their activities will remain. This, this new meaning of *Heuschrecke* is considered to be lexicalized. The metaphorical process which lead to a new reading of the word is evoked by an illusion to the the Bible, more precisely to the Old Testament, where swarms of locusts are sent as plague to mankind and their struggle to survive.

Example 3: „Szene“

This last example shall illustrate a somewhat more complex process on meaning change over a longer period of time. First, there is the WDG entry for this lexical unit:

Szene

1. Kleiner Teilabschnitt eines Theaterstücks, Films oder Hörspiels, Auftritt ['scene']

2. Schauplatz der Handlung eines Theaterstücks, Bühne [‘scene‘ (in a sense of place)]

3. Geschehen, Vorgang, Vorfall [‘incident‘]

Auseinandersetzung [‘conflict‘]

There is evidence in the Wortwarte corpus for most of these readings:

Ad 1. *Filmszene, Kussszene, Gerichtsfilmszene* (referring to a part of a film or theatre play)

Ad 2. *Szenenausleuchtung* (the illumination of e.g. the stage)

Ad 3. *Festnahmeszene* (a scene of detention)

Furthermore, the Wortwarte records many compounds where none of these readings of *Szene* would lead to an appropriate interpretation. We list the semantic classes of the words which combine with *Szene* in these examples:

Wettspielszene, Städtebauszene (betting, urban construction, i.e. activities)

Blechbläserszene, Intendantenszene, Friseurszene (musicians, theatre people, coiffeurs, i.e groups of people defined by a profession or shared interest)

The proecess of semantic change is a follows:

Szene ~= (1) a sequence of actions in a theatre play or film which is considered as a functional unit

➔ (2) The place, where these actions takes place (= a metonymic meaning extension, see *Szenenausleuchtung*)

➔ (3) a sequence of actions which is considered as a functional unit but which is not considered to be part of a theatre play of film but which reminds the viewer of a theatre play (metaphorical extension, see *Prügelszene* (fighting scene)).

➔ (4) A reference to the place where the aforementioned, non-theatrical activity (3) typically takes place (again an metonymic extension, see *Europaszene* (= Strasbourg, Brussels), *Berliner Szene*)

➔ (5) Extension from a small and limited sequence of activity to activities which have a long duration are permanent (*Entwicklungshilfeszene* (developmental aid), *Energieversorgungsszene* (energy supply)).

➔ (6)Reference to the (groups of) actors who are involved in such actitivites referred to by meaning (5), see *Blechbläserszene*.

The example of *Szene* is particularly appropriate to illustrate that a) the change of meaning can be a sequence of small steps, that b) aspects of the original meaning (e.g. reference to theatre) slowly fade and c) that different processes of meaning change can be linked to different aspects of the original meaning, in this example: place and people involved as well as the sequential character of activities (for a detailed description of processes of semantic change cf. Blank (1999) and Fritz (2005)).

The meaning of scene signifying a group of people which is characterized by a common profession or field of interest is now established. This can not only be proven by the high number of compounds in the Wortwarte which activate this meaning (~ 50 % of more than 700 compounds), but also by the fact that in the concept and word *Szenesprache* (~ in-group jargon) there is no further reference to a particular interest. The Duden publisher has recently launched an online dictionary “Wörterbuch der Szenesprachen” as a crowd-sourcing project aimed a collecting the up-to date in-grouop jargon (cf. Bathen et al. (2009) and www.szenesprachenwiki.de). *Szene* in this sense is also used occasionally as modifying part of a compound (e.g. *Szeneviertel*, in-group district in a city).

Conclusions and further work

This paper has presented three full-fledged example of semantic change which can be detected through the analysis of compounds in which the target word appears, in particular by semantically classifying the modifying part(s) of these compounds. The change of meaning could be illustrated with reference to a dictionary which faithfully records the state of the German lexicon in the 60s / 70s of the last century. Contrasting these well-established meanings with new compounds which have been collected in the past ten years revealed "gap" which have to be filled by stating now readings of the target words.

Two further questions arise with this approach which will be discussed in this section a) could the analysis be supported by means of natural language processing tools? and b) does the approach scale up to larger number of target words?

Could the analysis be supported by means of natural language processing tools?

The data analysis is preceded by some preparatory steps. These steps include the lemmatization of text words and the segmentation of compounds. There are tools available for solving this task (with German input words) at an acceptable degree of accuracy (e.g. the TAGH morphology, cf. Geyken/Hanneforth (2006) or the SMOR morphological analyser, cf. Schmid et al. (2004)).

The core work is that of analyzing the appropriate meanings of the compound parts (which can be more than two) and determine the semantic relation of these parts. The manual analysis revealed that domains of usage and broad semantic types (e.g. activity, product, group (of people)) play an important role in classifying the co-occurring compound parts. The main idea is that a lexicographer is alerted if there are groups of words which co-occur with the target words in compounds but are not semantically compatible with the known meanings of the target words. This would imply the existence of lexical resources which carefully and formally encode semantic fields, semantic types and selection preferences. Unfortunately, such resources are currently not available for German and on a broader scale. We can therefore currently delegate only the form based preparatory work to NLP tools.

A perspective for further automation of more advanced tasks might arise if processes of semantic clustering will gain maturity in the future. The author of this paper and his colleagues are working in this domain and hope to present promising results in the future, with impact on the task described in this paper.

Does the approach scale up?

The work which has been described in this paper has largely been done manually and has the character of a very limited feasibility study. The usefulness and the appropriateness of the approach had been at issue here.

The scalability of this approach to larger data sets hinges on two factors: a) fist of all, the meaning change has to leave "traces", at least it has to cause a change in the selectional preferences of the target word within compounds. In other words, the two meanings have to be sufficiently distinctive; b) the words which co-occur with the targe words in the targeted meaning have to be similar enough, e.g. they all have to belong to one and the same domain. Otherwise, clustering of these words would fail; c) the target word has to be a noun. Words of other parts of speech are by far not so productive as nouns are. E.g. for verbs and adjectives, it seems to be more promising to look at collocations.

Nevertheless, I hope to have brought home to important arguments with this studies:

1. I hope to have shown that nonce words are of much use in the context of broad-scale lexicological and lexicographical investigations; there has for long been a tendency to dismiss nonce words as completely uninteresting (cf. Kinne 1996);
2. I hope to have shown a way to establish a mechanism which could attract the attention of lexicologists and lexicographers to both individual meaning changes as well as large scale semantic changes in the vocabulary of a language. This work will, a.o., been integrated into the revision of the "Wörterbuch der deutschen Gegenwartssprache", which is part of the "Digitales Wörterbuch der deutschen Sprache" (DWDS. cf. Klein (2004)).

Lothar Lemnitzer

Berlin-Brandenurgische Akademie der Wissenschaften

lemnitzer@bbaw.de

Literatur

Bathen, Dirk / Sporer, Josefine /Deinert, Eva / Heiss, Martin (2009): *Duden – Das neue Wörterbuch der Szensprachen*. Verlag Bibliographisches Institut:Mannheim

de Beaugrande, Robert (1991): *Linguistic Theory: The Discourse of Fundamental Works. London:Longman* (also available at: http://www.beaugrande.com/ LINGTHERLinguistic%20Theory%20Title.htm, last checked: 21. July 2011)

Blank, Andreas (1999), "Why do new meanings occur? A cognitive typology of the motivations for lexical Semantic change", in Blank; Koch, Peter, *Historical Semantics and Cognition*, Berlin/New York: Mouton de Gruyter, p. 61–90

Fritz, Gerd (2005) .: *Historische Semantik*. Tübingen

Geyken, Alexander; Thomas Hanneforth (2006): TAGH: A complete morphology for German based on weighted finite state automata. In: *Proceedings of FSMNLP 2005*, Lecture Notes in Artificial Intelligence. Springer, vol. 4002, p. 55–66.

Herberg, Dieter / Kinne, Michael / Steffens, Doris: *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. de Gruyter, 2004 (also available online: www.owid.de, last checked: 21 July 2011)

Kinne, Michael (1996): Neologismus und Neologismenlexikographie im Deutschen. Zur Forschungsgeschichte und zur Terminologie, über Vorbilder und Aufgaben. In: *Deutsche Sprache* 24, (1996) 4, p. 327-358

Klein, Wolfgang (2004a). Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: Jürgen Scharnhorst (ed.): *Sprachkultur und Lexikographie.* Frankfurt am Main:Peter Lang, S. 281-309

Kramer, Undine (2011): Klappenbach, Ruth/Steinitz, Wolfgang (Hgg.) [1961-1977]. Wörterbuch der deutschen Gegenwartssprache (WDG). In: Haß, Ulrike (Hg.): *Große Lexika und Wörterbücher Europas. Europäische Enzyklopädien und Wörterbücher in historischen Porträts.* Berlin:de Gruyter

Lemnitzer, Lothar (2007): *Von Aldianer bis Zauselquote*. Tübingen:GNV.

Schmid, Helmut / Fitschen, Arne / Heid, Ulrich (2004): SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection, *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, p. 1263-1266.

Sinclair, J. (1982): Reflections on computer corpora in English language research. In *Computer corpora in English language research*, ed. Stig Johansson, p. 1-6. Bergen.

Widdows, Dominic (2004): *Geometry and Meaning*. CSLI publications, Stanford, California.