The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change

Daphné Kerremans, Susanne Stegmayr and Hans-Jörg Schmid

Abstract

Why do some new words manage to enter the lexicon and stay there while others drop out of use and are neither used nor heard anymore? Of interest to both lay people and linguists, this question has not been answered in an empirically convincing manner to date, mainly because systematic methods have not yet been found for spotting new words as soon as possible after their first occurrence and monitoring their early development and spread as exhaustively as possible. In this paper we present a new and improved tool which is designed to accomplish precisely these tasks when applied to material from the Internet. Following a brief review of existing tools for retrieving linguistic data from the Web (Section 2), we will introduce in some detail a tailor-made webcrawler, the so-called NeoCrawler, which identifies and retrieves neologisms from the Internet and stores data necessary for the systematic monitoring of their early development with regard to form and meaning as well as spread (Section 3). Following this description, we will present a case study discussing the results of an analysis of the neologism *detweet* with regard to its diffusion, institutionalization, lexicalization and lexical network-formation (Section 4). The study indicates that the NeoCrawler can indeed be applied fruitfully in the study of ongoing processes relating to how the meanings and forms of new words are negotiated in the speech community, how words spread in the early stages of their life cycles and how they begin to establish themselves in lexical and semantic networks.

1. Introduction

Which mechanisms are involved in lexical change and what language-internal factors (such as the morphological and phonological make-up of words) and language-external factors (such as the salience of the concept or referent and the authority of the coiner or early users) control these mechanisms? The methodological approach presented in this paper tries to tackle these long-standing and central questions in historical semantics by introducing a new method and by investigating – literally – new material, i.e. very recently coined neologisms. A neologism is defined here as a recently coined word[1] which is new to the majority of the members of the speech community. Unlike nonce-formations[2], however, neologisms are used with recurrent frequency, but are nevertheless still rare enough not to have become fixed and stable elements of the language.

While it may seem strange to look at new words in order to investigate historical change, the study of new words has a number of crucial advantages. Firstly, probably the most prominent asset – especially if one focuses on material retrieved from the Internet, as we do – lies in the possibility of collecting a more or less exhaustive sample of all authentic tokens of a new form within a certain period of time subsequent to its coinage. Secondly, the monitoring of recently coined words gives us the unique opportunity to study processes of ongoing change so to speak 'in vitro'. While lexicological theory has made a large number of claims concerning the early development of new words (cf. e.g. Bauer 1983: 42–61, Schmid 2011: 69–83), to the best of our knowledge these have never been tested empirically and systematically[3]. Is it true that meanings oscillate for a while and tend to rely on the context and co-text before they begin to stabilize? Is it true that forms are subject to variation before the speech community begins to agree on spelling, hyphenation and other formal properties? Is it true that changes in form and meaning (*lexicalization*) tend to go hand in hand with an increase in frequency of usage (*diffusion*)

---

1. Strictly speaking, the term *lexical unit* would be more appropriate here than the vague term *word*, since lexical innovations can concern various aspects of new linguistic signs. As such, a novel lexical unit can arise because both form and meaning are new, but also because a new form is paired with an existing meaning (very often for creative or pragmatic purposes) and vice versa (the traditional polysemy case). Tournier (1985) distinguishes between morpho-semantic, morphological and semantic neologisms. Since this paper deals exclusively with new words, i.e. new forms with new meanings, we have used the general terms *new word*, *new lexeme* and *neologism*, all of which are treated as being semantically interchangeable here.
2. See Hohenhaus (1996) and Stekauer (2002) for a detailed overview of nonce-formations.
3. Hohenhaus (2006) studies the diffusion process of the noun *bouncebackability* on the Internet, but does not consider other aspects of the lexicalization and institutionalization process. More recently, Buchstaller et al. (2010) use Google newsgroups to investigate a grammatical innovation, i.e. the decline and narrowing of usage of quotative *all* in favour of quotative *like*.

and the spread of words within the speech community, across text-types, registers and discourse domains and functions (*institutionalization*)?

The web-based methodology described in this paper aims to provide the means for answering questions of precisely this type. Before we embark on this endeavour, we would like to emphasize that we are well aware of the limitations involved in using only data from the Internet rather than 'real-life' texts and conversations. To an extent this limitation, which could only be overcome by means of very costly field work, is mitigated by the fact that many of the words we study are indeed 'born' on the Internet and are mainly used and spread there as well. And since the Internet plays an increasingly important role in the lives of an ever-growing number of people and is becoming more and more interactive[4], the general mechanisms and principles of new-word developments may not be too different from what goes on outside the Web after all.

This paper is a report on an undertaking which is very much in its infancy, as are the words it aims to investigate. It is therefore important to point out that the 'answers' suggested to the questions raised above are somewhat preliminary and will have to be investigated in future work.

2. Linguistic approaches to dynamic web-crawling

With an estimated 13.7 billion pages and an indefinite number of words (see www.worldwidewebsize.com)[5], the Web offers an amount and variety of language material that corpora cannot compete with. Even the currently largest corpus, the *Oxford English Corpus* (*OEC*), contains 'only' two billion words. Despite their careful compilation regarding text types

4. Even though the myth of the doubling of Internet traffic every three months has been proven wrong (Odlyzko 2003), the percentage of Internet users is still increasing steadily worldwide (Andrés, Cuberes, Diouf and Serebrisky 2007).
5. World Wide Web Size is a homepage run by Maurice de Kunder, who developed a method for estimating the size of the Surface Web (cf. de Kunder 2007). This figure, updated on a more or less daily basis, is based on the average of the indexes of Google, Bing, Yahoo Search and Ask, from which the amount of overlap between these search engines is detracted (cf. Gulli and Signorini 2005). The size of the index in turn is calculated through a daily query of 50 words extracted from a one-million-word corpus following Zipf's Law. In order to calculate the size of the search engine's index, the number of returned pages is multiplied by the relative frequency of the word in the corpus.

as well as social, regional and stylistic varieties, corpora remain static snapshots of the language at a given time. Corpora using the Web for their language make-up, such as ukWaC or the *OEC*, are also affected by this temporal rigidity, despite regular updates. While in principle language change can be studied with the help of comparable static corpora representing different synchronic cross-sections of a language (see e.g. Mair 2006, Leech et al. 2009), for the purpose of neologism-monitoring the time lag between data collection and public access is a crucial problem. This is also true for continuously augmented corpora such as the *Bank of English*, which are also known as *monitor corpora* (cf. McEnery, Xiao and Tono 2006: 67–69), because words that are new at the time of corpus compilation tend to be either obsolete or firmly lexicalized and institutionalized by the time the corpus is available for research. As a result a detailed investigation of these processes has become impossible or can only be carried out in hindsight and with great difficulty. Therefore, the timely discovery of potential candidates is of utmost importance for the study of the processes going on in the early phases of the establishment of neologisms. Before we introduce the NeoCrawler, we will briefly discuss two types of existing crawling approaches in linguistics: downloadable crawlers, which are not available for online use on the Net, but are installed on and operated from a desktop computer (Section 2.2), and on-demand crawlers accessible online (Section 2.3).

2.1. Downloadable crawlers

2.1.1. KWiCFinder

Like the NeoCrawler, KWiCFinder (cf. http://kwicfinder.com/) uses a commercial search engine to access the Web and generate user-defined language material. Queries are submitted to AltaVista, downloaded as HTML or .txt, summarized and documented with KWiC display. In addition, users also have the option to search the Web with the Java application WebKWiC, which retrieves cached website copies from Google and is considered to be more user-friendly by the developer (cf. Fletcher 2007: 36). Special search features include enhanced wildcard and "tamecard" options (Fletcher 2007: 34), which yield syntactic and orthographic alternatives for any given word. Queries can be expanded or narrowed down by means of "inclusion and exclusion" criteria (Fletcher 2001: 34), restriction searches to specific words, pages, dates and hosts, which are entered together with the search string. Post-processing tools include conversion into XML format as well as annotation and classification options. Unfortunately, Fletcher remains rather vague in this respect.

2.1.2. GlossaNet 2

Unlike KWiCFinder, GlossaNet 2 (cf. http://glossa.fltr.ucl.ac.be/) uses RSS and Atom feeds[6] to collect linguistic data. The original GlossaNet of 1998 was restricted to newspaper texts. In both versions, the user selects predefined feeds or adds some of their own and compiles a corpus to which the query is submitted. These pages are crawled in regular intervals and added to the corpus via the so called "Manager" (Fairon, Macé and Naets 2008: 3). The Manager not only retrieves the feeds from the server, but also sends them to the next server, which will perform boilerplate stripping, i.e. removal of programming code and duplicates. The second server subsequently assembles the corpus and is responsible for tokenization, lemmatization and tagging. The final results are then returned to the Manager, which informs the user that their queries have been performed and the corpus has been created and/or updated. Despite creating a dynamic corpus, which would enable neologism researchers to keep track of chronological developments, GlossaNet 2's reliance on a selection of RSS and Atom feeds provides only very specific information within a fairly narrow range of genres and semantic domains.

2.2. On-demand crawlers

In contrast to the crawlers described above, on-demand crawlers are available on the Web, where any user can consult them whenever necessary.

2.2.1. Kilgarriff's Linguistic Search Engine

Kilgarriff's Linguistic Search Engine (LSE) consists of five components[7]. The first one, the web crawler, performs daily crawls and feeds them into the LSE database, which is updated once or twice a year. While this may be sufficient for all kinds of applications of LSE (cf. Kilgarriff 2003: 3), this restriction poses a serious problem for the systematic study of very recent neologisms. The second component is responsible for filtering and

6. RSS and Atom feeds are tools that enable users to update, publish and exchange web content easily. They contain basic information about the content, such as title, link, description and publication date in XML format. GlossaNet 2 uses this link to access and download the page into the corpus.
7. To our current knowledge, the LSE has not been realized (yet).

classifying the crawled results. All material that does not contain 'real'[8] sentences, such as images, sound, lists of prices and people, is removed. The remaining pages are converted into standard XML format and their language is automatically identified with a Unicode compliant classifier. Pages are classified according to parameters such as text type and semantic domain with the help of TypTex and TypWeb tools (Folch et al. 2000). After filtering and classification, the linguistic processor, supported by the IMS Corpus Workbench where possible[9], performs tagging, parsing and lemmatization. After completion of linguistic post-processing, the results are stored in a database. Subsequently, the statistical summarizer Word Sketch (cf. http://wasps.itri.bton.ac.uk/) can be used to create automatic summaries of a given word's behaviour.

2.2.2. WebCorp Linguistic Search Engine

The WebCorp Linguistic Search Engine represents an improved and expanded version of the 1998 WebCorp programme (Renouf 1998). The most important change is the development of an independent linguistic search engine to access the Web, because of the various problems caused by commercial search engines (see 3.2.2 below). The proposed independent linguistic search engine is currently limited to *The Guardian* and *The Independent* newspaper websites and works progressively, i.e. only results collected on the crawling day are fed into the corpus (cf. Renouf, Kehoe and Banerjee 2005: 8). The authors have developed a vast and impressive array of crawling and post-processing features, such as exclusion lists, requerying of failed pages, wildcard and POS search options, neologism detection and collocation extraction. Despite the enormous potential for linguistic research, the WebCorp Linguistic Search Engine is not yet available for public use.

At present the WebCorp version (http://www.webcorp.org.uk/) available on the Internet still operates with commercial search engines. Query options include case sensitivity, output format in HTML or plain text, the size of the concordance span, the number of pages to visit (500 maximum) and options to search specific domains only and include or exclude specific words. Before the results are displayed to the user, HTML code, banners,

8. Kilgarriff defines a sentence in terms of prototypical characteristics and suggests a heuristic formula to detect these in the flow of diverse language material on the Internet (cf. 2003: 3).
9. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 64)

links and ads are stripped and duplicates removed. In addition, the date, author, headline and subheadline of the page are automatically extracted. The user receives a list of all the tokens per page, highlighted in red, and has the option to visit the original page. A valuable feature is the error logging of failed pages: the user is able to see how many and which pages returned errors. Unfortunately, the results cannot be downloaded in any form, so that further linguistic analysis is complicated. Moreover, the results only remain available for 24 hours on the WebCorp homepage.

3. The Architecture of the NeoCrawler

3.1. Overview of the architecture

While the crawlers and linguistic search engines discussed in the previous sections are very valuable and sophisticated tools for the study of language material culled from the Web, none of them is ideally suited to supplying the kind of data needed for answering the questions posed in the introduction. The NeoCrawler, which tries to improve this situation, was initially developed to replace a downloadable crawler used in our first tests. At that time our focus was on observing a selection of neologisms, so the crawler's first module, the *Observer* (see 3.3), was designed to serve this purpose. Because of the extendable architecture, which relied on a database (see 3.2), the second module, the *Discoverer* (see 3.3), integrated seamlessly into the existing project.

In order to explain the mechanisms behind the web interface of the NeoCrawler, we will give an overview of the basic structure first. The figure below outlines the main tasks of the two central modules.

Module I, the Discoverer, attempts to detect new words on the whole Web as closely to their date of coinage as possible. Since the module is comparatively young and still in its testing phase, we will confine ourselves for now to crawling the latest blogs from Google Blog Search[10] in the first step (a). The NeoCrawler retrieves a list of the blogs offered for all of Google's categories (see Section 3.4) (b) and follows the hyperlinks to obtain the contents of the blog pages (c). The pages are stripped to plain

10. http://blogsearch.google.com

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 65)

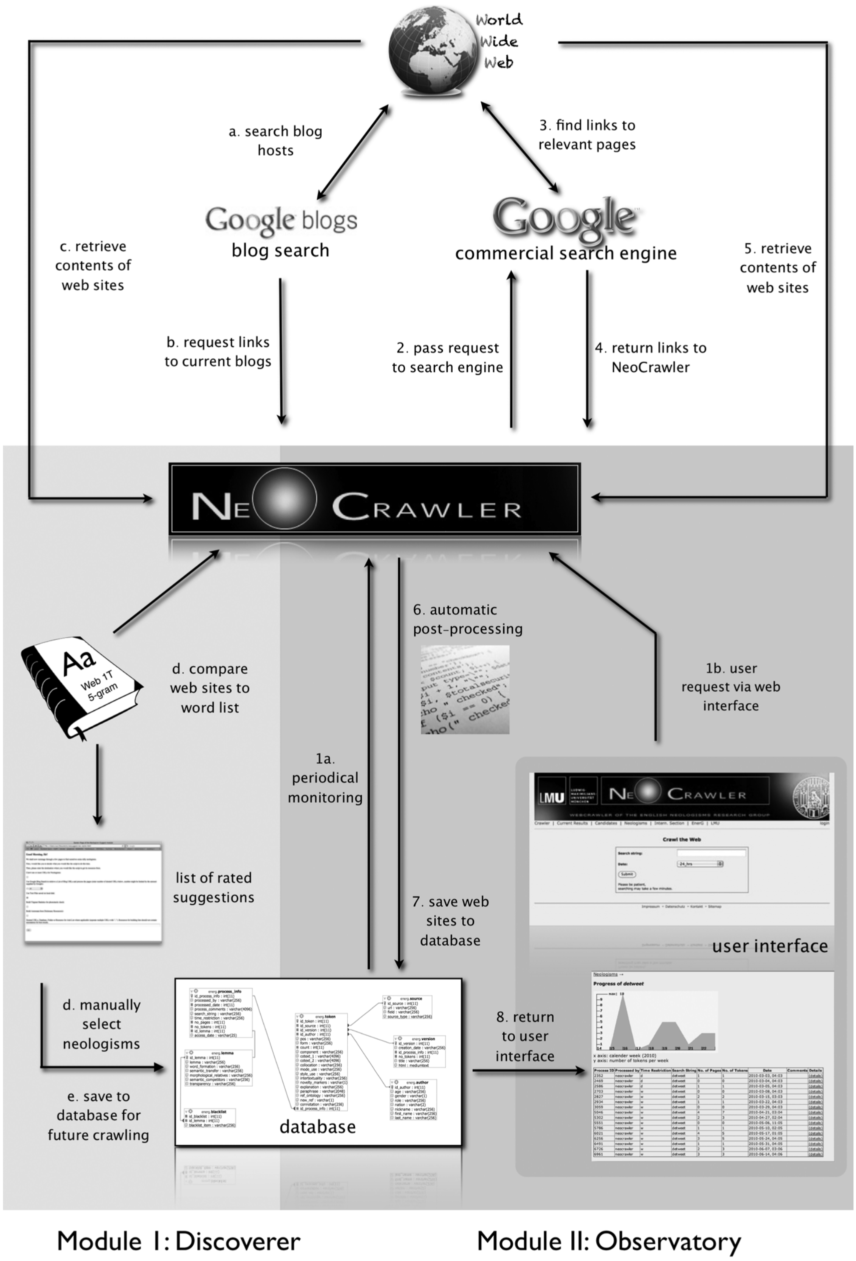
World
Wide
Web
a. search blog hosts
3. find links to relevant pages
Google blogs
blog search
Google
commercial search engine
c. retrieve contents of web sites
5. retrieve contents of web sites
b. request links to current blogs
2. pass request to search engine
4. return links to NeoCrawler
NeoCrawler
6. automatic post-processing
Aa
Web 1T
5-gram
d. compare web sites to word list
1b. user request via web interface
1a. periodical monitoring
list of rated suggestions
7. save web sites to database
user interface
d. manually select neologisms
8. return to user interface
e. save to database for future crawling
database
Module I: Discoverer
Module II: Observatory

Figure 1.

text and split into single words; then each word is compared to a previously compiled dictionary (cf. 3.4) to detect possibly unknown words (d). The words are subsequently analyzed with a *trigram filter* that compares the sequence of letters in the potential neologism with known typical patterns and rates the potential neologisms accordingly. The Discoverer then outputs a rated list of unknown words to the user interface in the web browser (e). The automatically generated suggestions have to be reviewed manually (f). Researchers can use the web interface to easily select the neologisms to be added to a database of neologisms (g), which will be crawled automatically in the future by the crawler's second module.

Module II, the Observer, handles the periodical searches for selected neologisms (1a), provides a public interface to the NeoCrawler (1b), and semi-automatically classifies the results. For the periodical observations, the NeoCrawler conducts a search for each neologism in the database. It compiles a web address with the search string and other parameters for Google, and passes the request to the search engine (2). Google treats the query like any other search process and searches the Web for relevant pages (3). The addresses of these pages are then returned to the NeoCrawler (4), which in turn follows each address and retrieves the contents of the pages from the Web (5). In the next step, the NeoCrawler partitions each web page to prepare its contents for the database (6). Both the entire HTML file and the automatically analyzed content of the search results are saved to the database (see 3.2) (7). From there, the data is passed to the web interface of the NeoCrawler (8), where the search results are permanently available to the researchers.[11]

The user interface offers various representations of the data, ranging from an outline of the diffusion progress of a neologism to basic statistics, detailed linguistic information and concordance lines. The data can also be downloaded in different formats, HTML and plain text, as well as in chronological order or classified structure to import the results in a concordancer, for example. With this survey in mind we will now have a closer look at the individual modules, beginning with the foundation of the NeoCrawler, its database.

11. Due to the restrictions imposed by Google's University Research Program (http://research.google.com/university/search/terms.html), the data obtained by the Observer is only accessible to our own researchers for the time being.

3.2. The database: Laying the foundation

Since it is common[12] in present-day corpus linguistics to annotate texts using the XML format[13], some explanation of why a database approach was chosen for this project may be required. The main reason is that despite its flexibility, XML is subject to a number of restrictions that make it insufficient for demands more complex than mere descriptive tasks. Basically, the structure of XML files is designed in such a way as to facilitate the hierarchical categorizing of (textual) data. Each unit or element, from page to morpheme level, is tagged, and the tags can be extended by any number of specifications. This facilitates very profound descriptions and in principle offers an unlimited number of markup options. The hierarchical structure offers many possibilities for single-user desktop utilization (cf. Carletta 2005).

However, it is this very freedom in manual editing that allows for the danger of inconsistencies in categorization and labelling, which make documents prone to errors in automatic processing. As a result, the file format has considerable drawbacks for the kind of large-scale server data mining required in this project. For example, the fact that it is virtually impossible to process complex computations with a large amount of data in the XML format has proven problematical. Processing XML files is slower in general, especially when it comes to searching and filtering, both central requirements for all kinds of data retrieval. In addition, complex relations in the source material need to be converted into the simpler hierarchical structure, which results in loss of expressiveness, unnecessary complication of data structures or redundancy of data. This either imposes restrictions on later analyses or requires duplication of data, especially when errors in the raw data have to be corrected.

A common alternative, which was chosen for the NeoCrawler project, is to store structured data in a relational database like MySQL or PostgreSQL. A relational database consists of a number of tables, each comprising columns with unambiguous headlines, and rows with the actual data (see figure 2).

12. Among others, Eckart (2008), Ide et al. (2002) and Dipper (2005) outline the methods of XML-based corpus annotation.
13. The Extensible Markup Language (XML) is specified by the World Wide Web Consortium (W3C, http://www.w3.org/XML/)

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 68)

process	
id_process	date
1	20100101
2	20100102

page		
id_page	name	id_process
1	I	1
2	II	1
3	1	2

token			
id_token	id_page	type	token
1	1	ART	The
2	1	ADV	quick
3	1	ADJ	brown
4	1	NOU	fox

Figure 2.

The rows of the tables are identifiable with a unique ID, which can be referred to in other tables as well. In a relational database, the smallest unit, such as a single token of a crawled neologism, is linked to rows of tables with more general information, for example the web page and its author(s). Thus, indirectly, the single tokens carry all the information available for them. The key feature of relational databases is that fields are linked, so any token can be tied up with any number of other tables. The advantage of this network of relations, unlimited in principle (compared to the hierarchical structure of the XML format), lies in the possibility of modelling facts of unlimited complexity.

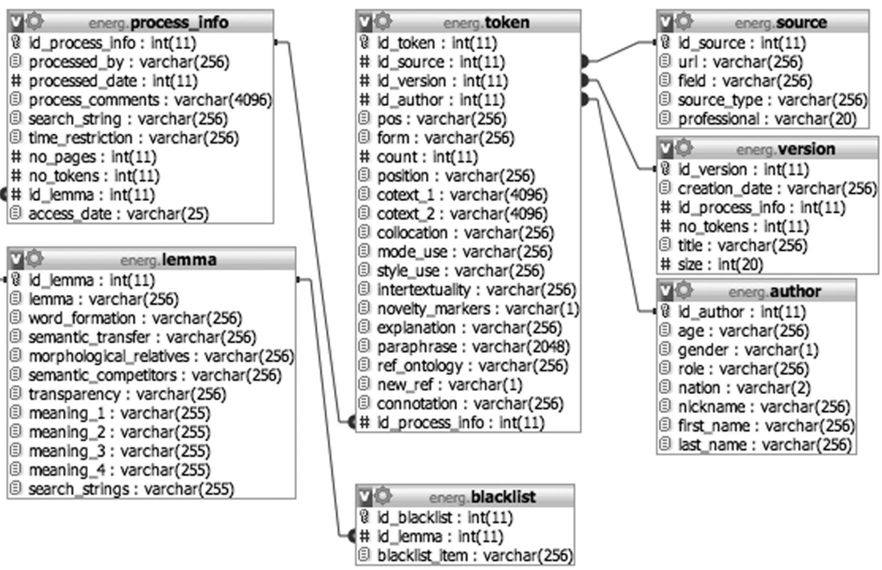

Figure 3.

In the case of our periodical observations (cf. step 1a in Figure 1), the database behind the web interface of NeoCrawler is modelled in exactly this way (see outline in Figure 3) and serves as both the source for the queries and destination for the results. Once a neologism has been added to the database for regular observation (table "lemma"), NeoCrawler gets the list of neologisms and initiates a search for each one. After the search process (see 3.) is completed, categorized information on the search results is stored in the database on four levels: process, lemma, page and token.

The headlines of the boxes in the figure above provide labels for their contents:

– The table "process_info" saves information retrieved for a given neologism in one crawling session, with one session corresponding to one 'process' uniquely identifiable and stored in the database. On the process level, the total number of pages and tokens found for the respective neologism are stored along with the date, the time restriction set in the query and the search string.
– As can be seen, the table "process_info" is linked to table "lemma", i.e. the *type* level: here information pertaining to the neologism can be specified and stored, e.g. the word-formation pattern, types of semantic transfer, such as metaphor and metonymy, semantic competitors (e.g. *google-cooking* as a competitor for *fridge-googling*), and, last but not least, meanings. This information has to be entered manually.
– Information on the page level is represented in the tables "source" and "version", which contain details about the web pages (see Section 3.3.3) that are retrieved in a search process, as well as their possible versions. Information on authors is specified in the table "author".
– Every single token of a neologism identified by the NeoCrawler receives one row in the table "token", containing a large number of cells including a co-text of 1000 characters and many other features such as the part of speech or the mode and style of use (see Section 3.3.3).

The connecting lines between the boxes point out the links to other tables and levels, which are represented by IDs (e.g. "id_source", "id_version", "id_author" in table "token") in the table rows. The last table, "blacklist", contains lists of strings that are to be excluded from the search results when crawling. The blacklist is the only table that is not directly linked to the "token" table, but connected with the lemmata instead, because its content applies to all results found for a lemma.

The principle of inter-linked tables containing information of increasing specificity avoids redundancy, which in turn enables complex queries and fast access to a large amount of data. With the linked data, the NeoCrawler is prepared for virtually any representation of the data and any kind of query, even though only basic computations are performed at present. The data does not need to be modified for more complex statistics, and server-based usage makes it possible for multiple researchers to edit the data simultaneously, even while the NeoCrawler is adding more results in the background.

3.3. The Observer: Monitoring neologisms

While in principle the Discoverer is of course the more basic module, as it identifies neologisms, we will nevertheless begin by describing the Observer, because some of its principles also provide the foundation for the Discoverer. Basically, the Observer contributes three crucial steps to the systematic acquisition of data on neologisms and their further processing for linguistic analysis: the web search, linguistic post-processing and classification.

3.3.1. Web search

The NeoCrawler uses Google to search for neologisms by means of an automated version of the same processes carried out in 'normal' manual Google searches. In a normal search scenario, a user enters a search string into Google's standard web interface, optionally adds a number of parameters such as date and language, and receives Google's response web page with a list of matching links. In responding to such queries, the Google Search web interface has the web browser encode the parameters set by the user. Following the user's click on the "submit" button, the web browser encodes a web address, also known as *uniform resource locator* (URL), with the search details. For example, typing the string "detweet" in the Google search form and opting for "100 results", "English" and "past week" in the advanced search menu will result in the creation of an URL like this (represented in slightly simplified form):[14]

http://www.google.com/search?q=detweet&num=100&hl=en&tbs=qdr:w&start=100

The parameters included in the search are more or less recognizable in this code, following abbreviations such as "q", "num", "hl" and "tbs". As an answer to the web browser sending this address, the search engine

14. For details see
http://yoast.com/wp-content/uploads/2007/07/google-url-parameters.pdf

compiles an HTML web page containing links to pages that match the selected criteria. All common web browsers display this HTML file as the well-known Google results page.

Rather than using Google's main search page manually, the NeoCrawler assembles the URL codes with all specified parameters itself, and fetches Google's answer by pretending to be a web browser. Since the periodical searches are carried out by the server at weekly intervals, the time parameter is currently set to one week, which ensures seamless retrieval of data more or less at the time they enter the Internet. Since 100 is the maximum number of results that Google returns for each call, the NeoCrawler requests a series of result pages for each neologism by varying the "start" value.[15]

Each HTML page returned by the Google server is then parsed by the NeoCrawler. It extracts all web links from it, i.e. links to pages containing the search string, and filters out Google-internal tracking links, blacklisted sites (see 3.2.1) and Google cache links. In this way, outdated and duplicate versions of websites are prevented from spamming the database, and the search process is kept as efficient as possible. In the next step, the NeoCrawler follows all remaining links from the search results and downloads the exact contents of the page, excluding pictures.

While the use of a commercial web engine like Google is not uncontroversial (cf. Kilgarriff 2003, Renouf, Kehoe and Banerjee 2005)[16], it can be argued in favour of this decision that Google allegedly has the largest number of indexed pages (cf. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html). Moreover, the index is updated fastest in comparison to other search engines, for many pages even on a daily basis (cf. Lewandowski 2008a: 820). As a result, Google shows the latest, updated versions of pages and is the leader in "freshness" regarding its

15. It should be noted that the NeoCrawler used Google's standard search interface in the pilot phase, which has a limited query rate. In the meantime, our project has been accepted by Google's "University Research Program for Google Search" (http://research.google.com/university/search/), which gives us the permission to run automatic queries with full access to Google repository.
16. The main criticism concerns the commercial ranking of results. As a result, statistical analyses are distorted, because the displayed pages might not accurately reflect the real use of a lexeme. Secondly, the absence of a wildcard search restricts the researcher's query options, but this can be solved by incorporating a search engine like Yahoo, with which such searches are possible. The problematic display of a limited co-text on the Google interface has been solved by setting the NeoCrawler's co-text extraction to 500 characters.

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 72)

index (Lewandowski 2008a: 824). Lewandowski furthermore investigated display delay and found that it is Google that again shows the lowest delay margin, 2 days on average, between the retrieval of updated pages and their inclusion in the Google search engine (cf. 2008a: 823). Fast discovery of new pages and re-retrieval of updates is qualitatively important, because research has shown that although the majority of pages change only marginally, approximately 8% of the web consists of new pages that go online every week and 20% of all web pages vanish within a year of their publication (cf. Ntoulas, Cho and Olson 2004: 3). Since Google scores best on quantity (the amount of indexed pages), quality (their freshness) and speed (both concerning retrieval and re-retrieval of updates), our current reliance on Google for web access appears justifiable.

3.3.2. Post-processing Features

When a web page has been retrieved and the full HTML version has been stored in the database, the NeoCrawler performs a number of automated analyses on the individual pages. It features further filters, syntactic parsing and suggestions for subsequent manual evaluation.

As users of Google know, Google's harvest tends to be quite confusing. Often a large number of potential hits turn out to be either false positives, i.e. pages that do not feature the string searched for (which is usually due to the fact that pages indexed by Google have been changed since indexing), duplicate copies, or otherwise useless pages. To increase the integrity and validity of the collected material, the NeoCrawler therefore checks each page for false positives and identifies exact duplicates or nearly similar versions of the same page with no relevant changes. Both types of page are removed from the list of pages prepared for parsing. Duplicates are reliably detected by comparing the title and the file size to all previous results of the same search. The NeoCrawler ignores the invalid pages in all subsequent computations and does not store their contents in order to keep both the database and the final output slim, but stores the addresses to ensure gapless coverage.

Subsequently, the remaining pages are stripped of all content irrelevant for linguistic analysis, such as HTML tags and script code. The result is the human-readable content of the web pages that can be displayed in any text editor and can be passed on for further linguistic processing to a concordancer, for example. Nevertheless, the complete page is still available in the database and can be viewed and downloaded in its original form at any time.

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 73)

Some results pages are not useful because their content is either encrypted or a mere compilation of links to other pages without linguistically valuable content. Facebook, for example, allows Google to search the content of the private member sites and returns their links, but their body is only readable for users logged in with a Facebook account[17]. Because of this, the NeoCrawler allows researchers to individually blacklist sites for the neologisms. Blacklisted sites will no longer be displayed in the current search results or previous ones, but they are kept in the database.

The next steps in preparing the pages for linguistic analysis relate to the content level. First, the NeoCrawler extracts the title of the document, breaks up the stripped content into words and sentences and identifies the relevant tokens, that is, the instances of the requested neologism. This is the process of tokenization. For each token found, the NeoCrawler saves a co-text of 500 characters around the target word, which can be used later for fully searchable concordance lines. The NeoCrawler also counts the number of tokens found on each page, adds up the number found on all pages of the corresponding search process, and stores the information in the database. With this information, the NeoCrawler can provide basic statistical data such as the page/token ratio. The second step is part of speech tagging. The stripped contents are automatically analyzed with an open source part-of-speech tagger[18], which considerably facilitates later analyses, e.g. concerning the collocational behaviour of the new words. Last but not least, the NeoCrawler detects novelty markers (e.g. *so-called*, quotes etc.), and adds information about them to the token table of the database.

3.3.3. Linguistic Classification

After post-processing, the pages are available in a form that linguists can use for further research. If the aim is to investigate the behaviour and development of new words from a language-internal and language-external perspective, as suggested in the introduction to this paper, one has to set up a classificatory system which captures not only their formal, morphological and semantic properties, but also textual and socio-pragmatic characteristics of their environment. The establishment of such

17. As a result, only publically accessible Facebook pages are included.
18. The Standford Log-linear Part-Of-Speech Tagger is licensed under the GNU General Public License (http://www.gnu.org/licenses/gpl.html) and can be downloaded free of charge from http://nlp.stanford.edu/software/tagger.shtml.

a framework is not entirely unproblematic, because the research undertaken in the field of computer-mediated discourse (CMD) has not yielded any reliable classification schemes for Internet text-types and genres, while categories established in traditional discourse analysis and stylistics (cf. e.g. Wehrlich 1976, Beaugrande and Dressler 1981, Biber 1988, 1989, 1995, 2007) are largely inadequate for capturing the variability, dynamicity and fuzziness of the material found on the Internet.

Biber (2007: 116), for example, proposes the four text-type dimensions "personal, involved narration", "persuasive/argumentative discourse", "addressee-focused discourse" and "abstract/technical discourse" on the basis of statistical multi-dimensional analysis, which uses text type-specific linguistic features. However, suitable as this framework may be for "traditional" texts, these four types seem to be too broad to reflect the range of variation found on the web.

An approach which comes closer to meeting the demands of this project is Herring's "faceted classification scheme" (2007), which adapts Dell Hymes' (1974) SPEAKING model to CMD. Herring argues that the various CMD forms are the result of interaction between technological and situational influence factors, which she calls "facets" (2007: 10). Both facets are open-ended and dynamic. Social-situational facets include topic, purpose, tone of the message as well as structure and characteristics of the participants. The technological dimension captures several medium factors such as synchronicity or 1-way vs. 2-way message transmission. This dimension is indeed very important for linguistic issues, because technological innovations have created new forms of communication, e.g. Twitter, and are of utmost importance in the diffusion process of neologisms.

Since Herring's system is too fine and detailed to be applied for the present purpose where thousands of pages await linguistic classification, we have taken it as an inspiration for a somewhat simpler two-level multi-dimensional[19] classification, which tries to balance practicability and adequacy (cf. Table 1).

A primary distinction at page level is made between meta- and object-linguistic modes of use. Since profuse talking about, rather than referential use of, a new lexeme is assumed to inhibit lexicalization (cf. Metcalf 2002: 155–157), we first identify those instances that merely define, paraphrase

19. We do not use dimension in the sense intended by Biber (2007) as synonym for text types. In our approach, the dimensions represent linguistic perspectives on classification.

Table 1. Page-level classification scheme

Mode of use	**Metalinguistic**	
	Object-linguistic	
Semantic features*	**Field of Discourse**	**Sub-field of Discourse**
	general	
	politics	
	law	
	business	
	sports	
	science	
	advertising	
	lifestyle	celebrities, food and drink, fashion, health, other
	entertainment	radio and TV, movie, music, other
	computing/Internet	gaming, technology, business, other
	other	
Socio-pragmatic features	**Type of Source**	**Sub-type of Source**
	Blog	
	News	
	Discussion groups	
	Portal	directory, jobs, community, Hollers, Gather, Bebo, Blippy, other
	Social Networks	Facebook (public), MySpace, Meetup, other
	Filesharing	documents, music, video, photo, blog
	Microblogging	Twitter, Tumblr, other
	Self-reference	
	Academic	
	Dictionary and thesaurus**	
	Other	
	Authorship	
	Private	
	Professional	

* not applicable to metalinguistic uses
** only applies to metalinguistic uses

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 76)

or comment on the given neologism. The top level furthermore involves the dimensions "field of discourse" and "type of source", both of which are more or less explained by the categories listed in Table 1. A third dimension is concerned with authorship and only applied to a small number of categories. Certain types of discourse contain an inherent authorship status: the people who write for established newspapers will be professional journalists, but the majority of discussion group users will use the forum for personal reasons. Blogs, however, can fulfil both functions: on the one hand they replace the old-fashioned diary or internal monologue, and on the other hand, they are used by professionals as an extension of or a complement to their work. We therefore distinguish between private and professional authorship. Although the distinction between private and professional blogs is not always straightforward, several linguistic and visual differences set them apart from each other. In professional blogs, for instance, a lot of space is filled with advertisements, much more so than in private blogs. Furthermore, professional blogs more frequently name the author or use the generic admin, whereas private blogs are characterised by authors publishing under nicknames or pseudonyms. Unfortunately, the geographic origin of a page[20] does not necessarily correspond to the current location of a user, let alone to his or her background. For some pages only, regions can be determined manually by relying on the information users share, for instance in discussion groups or blogs. The location of the author is thus deemed too unreliable to be included as a variable.

The lower classification level is concerned with a linguistic description of the individual tokens. Whereas we have assumed semantic, socio-pragmatic and to a certain extent also textual homogeneity on the page level, the different tokens contained on a single page might differ with regard to a range of linguistic properties. Table 2 shows the classification scheme on the token level, which contains categories that are all more or less well established in linguistic terminology.

At present, classification proceeds manually, assisted by drop down menus on the interface of the Observer. This process is to be automatized as far as possible by means of URL parsing for the semantic and socio-pragmatic types and fields of discourse. Apart from automatic part-of-speech identification with parser and tagger, we aim to integrate further

20. The geographical location of a web server can be determined by the IP adress, a practice called *geolocation*.

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 77)

Table 2. Token-level classification scheme

Linguistic dimension	Class label	Class realization label
Syntax feature	Part of speech	verb, noun, adjective, adverb, interjection, phrase, other
Text feature	Position	banner, title, headline, body, footer, signature, caption, teaser, category, tag
Metalinguistic feature*	Explanation	definition, paraphrase, none
Sociolinguistic feature	Style of use	neutral, formal, informal, vulgar, e-speak
Cognitive feature	New referent	yes/no

* only applies to metalinguistic uses

tools that reduce the amount of manual classification required; a certain degree of manual labour will most likely remain indispensable.

3.4. The Discoverer: Identifying neologisms

Besides monitoring the development of known neologisms, one of the most important aims of the NeoCrawler project is to identify new words in the World Wide Web. Our vision is to find them on the very date of coinage and observe their development from that point on, but given the current size of the Internet – Google's index listed eight billion pages in 2005 (Uyar 2009) – and the complexity of web technologies in general, this is an ambitious aim which we can only approximate for now. The NeoCrawler rises to this challenge in two ways.

The first method tackles the task with the help of the Observer by targeting metalinguistic markers of linguistic novelty. This means that the NeoCrawler searches for strings such as

– came up/ made up/ with a/the (new) term/word
– invented a/the (new) term/word
– coined/ heard/ read / stumbled upon a/the (new) term/word.

The results output produced by the NeoCrawler is a table that displays the search strings in context along with the option to save a new word to the database for future observation. Once added to the database, the neologism will be automatically included in the upcoming and all future crawling rounds. In the list of results to be reviewed manually, however, only the search string such as "stumbled upon a new term" can be automatically identified within the web page and thus highlighted. As a consequence, the researcher has to read and analyze large parts of the co-text to detect a new word, which is a time-consuming procedure. Another obvious disadvantage concerns the time of detection. Since we are relying on pages where people already talk about a new word, we are always one step behind, even though first attestations of neologisms are usually found in the first search, which is always conducted without time restriction.

The second method, implemented more recently and referred to as *The Discoverer*, tries to reduce the time gap between coinage and identification by means of a direct automatic analysis of web pages. This also has the advantage of drastically decreasing the necessary amount of manual intervention. The Discoverer was programmed by René Mattern, to whom we are greatly indebted, as part of his M.A. thesis in computational linguistics. At the time of writing, the Discoverer is in its testing phase, in which it does not yet crawl the entire Web for neologisms. The Discoverer module is operated with a separate web interface that currently offers two possibilities: on request, the NeoCrawler searches for neologisms either in blogs on the Internet or in files on a local hard disk. In the case of the blog search, we have so far consulted only a few blogs preselected by Google on Google Blog Search[21]. For the time being, the blog search retrieves an individually specified number of blogs of all available categories[22].

In the next step, both blogs and files from the hard disk are prepared for processing. The downloaded HTML files are stripped of all linguistically irrelevant content such as HTML tags and programming code, date and time, email addresses and URLs, and the NeoCrawler extracts the body of the blogs. The files and the plain text of the blogs are then split into single words, using capital letters and punctuation marks as delimiters between words. The remaining words are compiled into a list sorted by frequency in the text. This list is then passed through a set of filters. In

21. http://blogspot.google.com. Admittedly, we are subject here to commercially motivated selection by Google, but we intend to detach from the search engine in the near future to extend our blog search to all major blog providers.
22. At the time of writing, Google Blog Search presents current blogs of the following categories: politics, US, world, business, technology, video games, science, entertainment, movies, television and sports.

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 79)

this process, the NeoCrawler eliminates stop words[23], words with fewer than three letters and words containing more than two digits. Proper names are filtered out by consulting a database of proper names contributed by a cooperating department[24]. All remaining words are then compared to a reference dictionary and a user-generated catalogue of known words, which is currently based on a reduced version of Google's web-scale N-grams[25]. The N-gram Corpus was created in 2006 and consists of about a trillion running words taken from web pages. This data is organized as unigrams, bigrams and so on up to five-grams. Taking into consideration the size of the corpus, we decided to use the approximately 14 million unigrams, i.e. single words, as a start, and also removed non-words according to the same criteria later applied to the blogs. The resulting dictionary still contains more than 7.8 million tokens, which helps the NeoCrawler to filter out most of the words used before 2006, as well as common typing errors and misspellings.

The general output of the Discoverer still contains many items which clearly are not new words, or in fact are not words at all. Therefore, it rates the remaining words by performing a trigram analysis on the sequences of letters. The NeoCrawler contains a database of trigrams (a sequence of three letters), which is a list of all three-letter substrings of Google's N-grams database and their respective frequencies. We assume that the trigrams represent typical sequences of letters in English words. With this reference, the frequencies of all trigrams within a potential neologism are used to calculate the probability that it is an English word. The words with the lowest values are dropped.

At this point, the number of potential neologisms per average web page is down to less than ten, and the researcher has to go through this list of candidates manually and decide for each word whether it is a neologism, a known word or not a word at all. The NeoCrawler saves all words marked as "known" and "not a word" (including typing errors and misspellings) in two user-generated catalogues, which augment the N-grams database, so they will be ignored in future analyses. With these growing catalogues, we hope to soon decrease manual intervention to a minimum.

---

23. Stop words are extremely common words that typically cause problems in natural language processing and are therefore typically extracted prior to natural language processing (Luhn 1958).
24. We are indebted to Michaela Geierhos and the Centrum für Informations- und Sprachverarbeitung, LMU München; cf. Geierhos (2007).
25. Google's N-grams are freely available at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

If a word is marked as a neologism, NeoCrawler saves it to the database. From then on, the Observer module will include it in the periodical crawling processes and analyze the results in the way described above.

4. Applied NeoCrawling: *detweet*

In the following section we present a case study which illustrates some of the NeoCrawler's functions and applications in the field of neologism-monitoring. Our focus will be on the practical aspects of monitoring the diffusion, lexicalization and institutionalization processes observable for the young lexeme *detweet*. The study is based on no more than 144 tokens of this form found up to April 2010 and of course cannot claim to come close to presenting statistically reliable analyses and interpretations. We have selected this small dataset for our case study because it provides maximum transparency for all stages of the application of the NeoCrawler.

The notion of *diffusion* is used to refer to the spread of a new word as measured in terms of discourse frequency, or more precisely in the present context, in terms of the number of tokens and types of new words found on Internet websites. *Institutionalization* is defined in a fairly narrow sense (as compared to, e.g. Bauer 1983: 48, Lipka 2002: 112, Brinton and Traugott 2005: 45–47) as a process of spread across text-types, register and genres, both within and outside the Internet, as well as across the fields of discourse mentioned in 3.3.3. The rationale behind this notion is that in addition to sheer frequency, the "success" of a new word is reflected in its spread across different socio-pragmatic situations and the purposes for which it is used. In line with existing suggestions (cf. e.g. Bauer 1983: 42–61, Brinton and Traugott 2005, Schmid 2011: 69 ff.), *lexicalization* is regarded as a cover term for structural changes undergone by neologisms, i.e. morphological, grammatical or semantic developments. *Conventionalization* will be used as a cover term subsuming *diffusion* and *institutionalization*, while *establishment* includes all three types of process.

4.1. First recorded occurrence

Detweet is one of the more recent coinages that have arisen after the introduction of the popular microblogging service Twitter. The sentence in (1) represents the first use that was found by the NeoCrawler in May 2008, when it appeared on a Question and Answer portal page called *AskMosio*. From a morphological perspective, *detweet* is the result of a prefixation process and consists of the prefix *de-* and the basis *tweet*, which is used as a noun and verb referring to 'a Twitter message' and 'to post messages on Twitter' respectively. Using the ablative prefix *de-*, *detweet* denotes the removal of Twitter messages or tweets, i.e. 'to delete a tweet'.

(1) Can you delete your twitters? yup, login to twitter.com, then select the trashcan by the tweet you want *detweeted*. (my 1000th answer!!!).

In spite of the fact that the meaning of *detweet* in (1) is fairly clearly 'delete', not all of the word's uses during its early stage of conventionalization allow for a similarly unambiguous semantic analysis. An occurrence of *detweet* in a tweet in October 2008, given in (2), poses a problem, for example. Although the co-text, which is restricted to 140 characters on Twitter, does not provide enough clues to assign a distinct meaning, it seems certain that the sense 'to delete' does not apply here:

(2) What is everyone going to do with their Twitter withdrawal time tonight? Is there a cure for the DT's (*DeTweets*)?

Judging from the preceding phrase *Twitter withdrawal time*, the prefix *de-* might be interpreted as a negation of *to tweet*, yielding 'not to tweet'. The presence of the definite article *the* however, excludes a reading as a verb and suggests that *DeTweet* functions as a noun. This not only shows that, as predicted by lexicological theory, the meanings of new words are variable and subject to modifications, but also that their grammatical status seems to stay flexible. This should be kept in mind when we now proceed to report on the early diffusion of the form *detweet* and its semantic development.

4.2. Diffusion

By April 2010, the NeoCrawler identified a total number of 117 web pages that contained *detweet* in one of its word forms. Table 3 shows the distribution of tokens grouped according to word classes and word forms. As the table shows, the majority of the 144 tokens extracted by the concordancing software CasualPConc[26] are verbal forms (130 tokens, constituting 90.2%). Within the verbal paradigm, base-form occurrences accounted

26. CasualPConc is a freeware concordancing programme for Mac OS X. It works similarly to other concordancers like AntConc, but includes the advantage of concordancing parallel corpora. CasualPConc can be downloaded from http://sites.google.com/site/casualconc/, together with other CasualConc tools.

Table 3. Tokens per word form ratio

	Verbal forms				Nominal forms		Total
	detweet	detweets	detweeting	detweeted	detweet	detweets	
Tokens	74	2	34	20	12	2	144

for the lion's share (56.9%), followed by the present participle form *detweeting* (26.1%). Although one of the first known uses, as illustrated in (2), was in nominal form, the token analysis suggests that *detweet* is spreading in the Internet/speech community mainly as a verb.

To provide an idea of how the diffusion of *detweet* has proceeded so far, Figure 4 represents the overall number of pages per month that contain this form and its variants.[27]

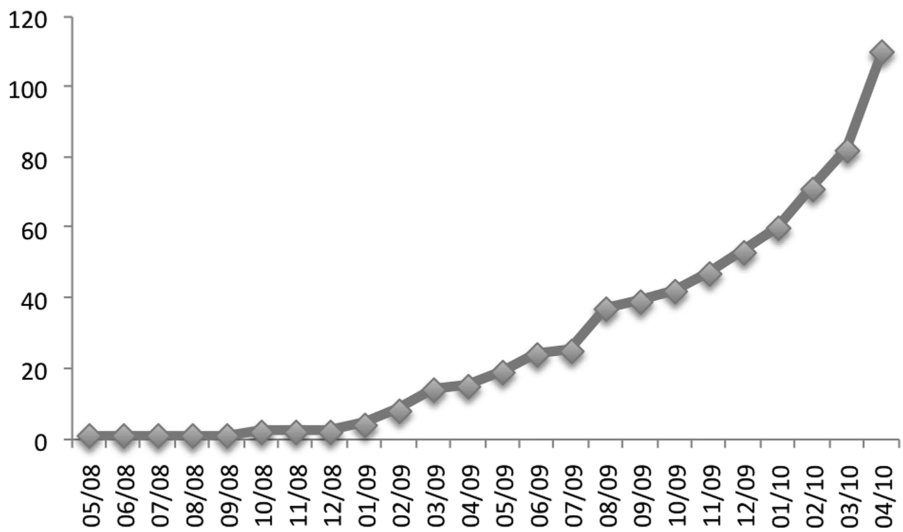

Figure 4. Cumulated pages per month

27. Seven pages whose publication date could not be traced have been omitted.

While the curve in Figure 4 suggests a continuous and constant increase in numbers of websites, this does not in fact do justice to the dynamics of the diffusion process. To provide a more detailed picture, Figure 5 charts the number of newly uploaded pages which were identified by the Neo-Crawler at weekly intervals in the period from the first attested use in May 2008 up to April 2010.

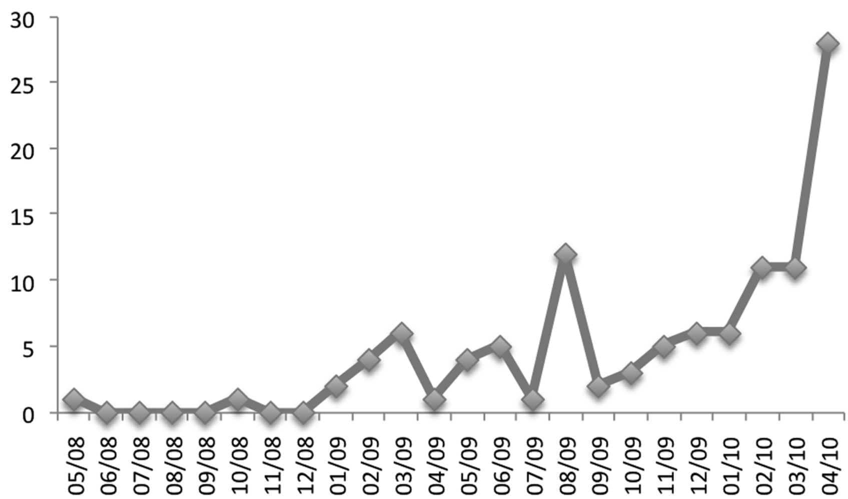

Figure 5. New pages per month

This figure indicates that rather than seeing a linear increase in the number of websites containing *detweet*, ups and downs can be observed, reflecting more or less intense communicative activity using the form *detweet*. Looking at Figure 5, the most striking peaks are found around August 2009 and in early 2010. Two extra-linguistic events appear to be responsible for the increased use of *detweet* in August 2009. Firstly, at that time, J.R. Smith, a well-known NBA player, decided to suspend his Twitter account in the wake of some controversial tweets which stylistically resembled the discourse of a certain street gang. The original article entitled "J.R. Smith decides to deTweet" appeared in the Denver Post[28] and was afterwards taken up in a specialized blog and discussion forum[29].

28. http://www.denverpost.com/nuggets/ci_12993784
29. http://www.binarybasketball.com/forums/threads/9718-J.R.-Smith-decides-to-deTweet

Almost simultaneously, the Twitter account of a somewhat dubious businessman was deleted by Twitter itself, because he had been trying to raise money for another one of his suspicious activities[30]. Although this news did not spark an article in any of the established newspapers, it was passed around in several community portals, among them *everyjoe.com* (31 August 2009):

(3) [. . .] In the End, Rawman Was *Detweeted*. (http://www.everyjoe.com/articles/franchise-founder-loses-twitter-food-fight/)

This example confirms our earlier observation that different meanings, 'to give up tweeting' in the J.R. Smith case and 'to be kicked out by Twitter', are competing with each other. In addition to (3), there are only two more uses in which the passive form *be detweeted* refers to the act of being removed from the Twitter service.

4.3. Lexicalization

As predicted by lexicological theory (cf. e.g. Lipka 2002: 110 ff.; Schmid 2011: 73–83), then, the recent coinage *detweet* still seems to be both grammatically and semantically – and, incidentally, orthographically – unstable, or, and this remains to be observed in the future, has already embarked on developing a system of polysemous senses associated with the form. In this section we will leave the level of the diffusion of the form in the (cyber-)speech community and move to a semantic investigation of the data reaped by the NeoCrawler.[31]

The most frequently used meaning in the data available so far, which can be rendered as 'sign off', is illustrated in a tweet from April 2010 in (4):

(4) *Detweeting* until 3–5 pm. If needed DM/text/email me.

This sense is instantiated in 29.5% of the records. What is important is that of the 36 tokens, only one is metalinguistic in nature, which indicates that this sense is currently the preferred 'normal', i.e. object-linguistic, use in the speech community. Since the denotatum is clearly an action, it is

30. http://www.everyjoe.com/articles/franchise-founder-loses-twitter-food-fight/
31. Eighteen pages, where the meaning could not be disambiguated or determined on the basis of the often insufficiently informative co-text, were omitted from further analysis.

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 85)

hardly surprising that *detweet* typically occurs as a verb (in the infinitive or as the present participle).

In example (5), the author explicitly explains his definition of the word *detweet* as signifying the opposite of the more well-known *retweet*[32]. Example (5) was taken from the author-coiner's blog post in February 2010. In contrast, the second most frequently found sense of *detweet*, is mostly used as a noun and in metalinguistic uses. *Detweet* in this sense of 'forwarding a tweet with disapproval' accounts for 23.7% of the tokens, but the majority of these occurrences are metalinguistic comments such as the definition in (5) or references to this blog entry. In example (5), the author explicitly explains his definition of the word *detweet* as signifying the opposite of the more well-known *retweet*[33]. Example (5) was taken from the author-coiner's blog post in February 2010.

(5) So I'm going to just De-Tweet it in the same way people Re-Tweet stuff. I hope to start a trend. The *DeTweet* Defined: DeTweet (AKA: De-Tweet or DT) = Passing along the tweet of another with some degree of disapproval. It can range from strong (that's a lie) to mild (there are exceptions or conditions).

Detweet in this sense of 'forwarding a tweet with disapproval' is the second most frequent usage.

The meaning evoked by (6), synonymous with 'to unfollow', i.e. to stop following someone's tweets, was identified in 17 tokens (13.9%). For this sense, only one metalinguistic result was recorded. Similarly to the first meaning 'sign off', the action-like character of the word is reflected in its exclusive use as a verb in the entire inflectional paradigm. Although the third person singular form was found only once, the other morphological options did not show any preferences. This particular meaning is illustrated in (6), which was found in a private blog post in March 2010.

(6) I mean Barack Obama, Martha Stewart, Dame Elizabeth (whom I had to *detweet* for spamming me about that whole Michael Jackson nonsense) never started following me.

Finally two usage-types can be identified which occur predominantly in passive mood. The first, 'be removed from Twitter' was already illustrated in example (3) above ("Rawman was *detweeted*"). In addition, the object

---

32. *To retweet* means 'to post a tweet of another user on your page, because it is funny, important, meaningful, etc.' It is followed by the abbreviation RT.

33. Footnote Missing?

of the *detweeting* process can also be a Twitter message deleted by the Twitter team, as demonstrated in example (7) from a private blog in March 2010.

(7) *Detweeted*. One of my tweets disappeared today. It wasn't a latency issue – sometimes text tweets to Twitter appear several hours later or never appear at all. This tweet was in my stream long enough to receive a reply and to be referenced in another tweet before it went missing. I didn't delete it, and I've never experienced or heard chatter about spontaneously combusting tweets before, which led me to wonder if Twitter administrators deleted it because they considered it offensive.

It could be argued that the sense in (7) is a semantic narrowing of 'to delete', as it is not the individual user that decides to remove their tweets, but the Twitter authorities. A mere 8% of the tokens are uses of this type. In terms of grammatical form, 8 out of 10 tokens were the past participle, once the third person plural form preceded by Twitter as subject was found. Meaning and grammatical form thus strongly correlate.

Table 4 provides a summary of the five senses identified in the dataset and cross-tabulates them with their grammatical distribution.

While it is impossible of course to predict if some or only one of the five meanings will eventually win the race for establishment and push out the others, or whether a system of five polysemous senses will stabilize, it is interesting to chart the temporal development of the senses. This is rendered in Figure 6 which gives the timeline of the frequencies for each of the five semantic usage types.

Table 4. Grammatical-semantic distribution per word form

	detweet (V)	**detweet (N)**	**detweets (V)**	**detweets (N)**	**detweeting**	**detweeted**	**Total**
1) to sign off	15				19	1	35
2) to delete	22		1		2	5	30
3) to pass along with disapproval	16	6			3		25
4) to unfollow	7		1		4	5	17
5) to be removed from Twitter	1				1	8	10

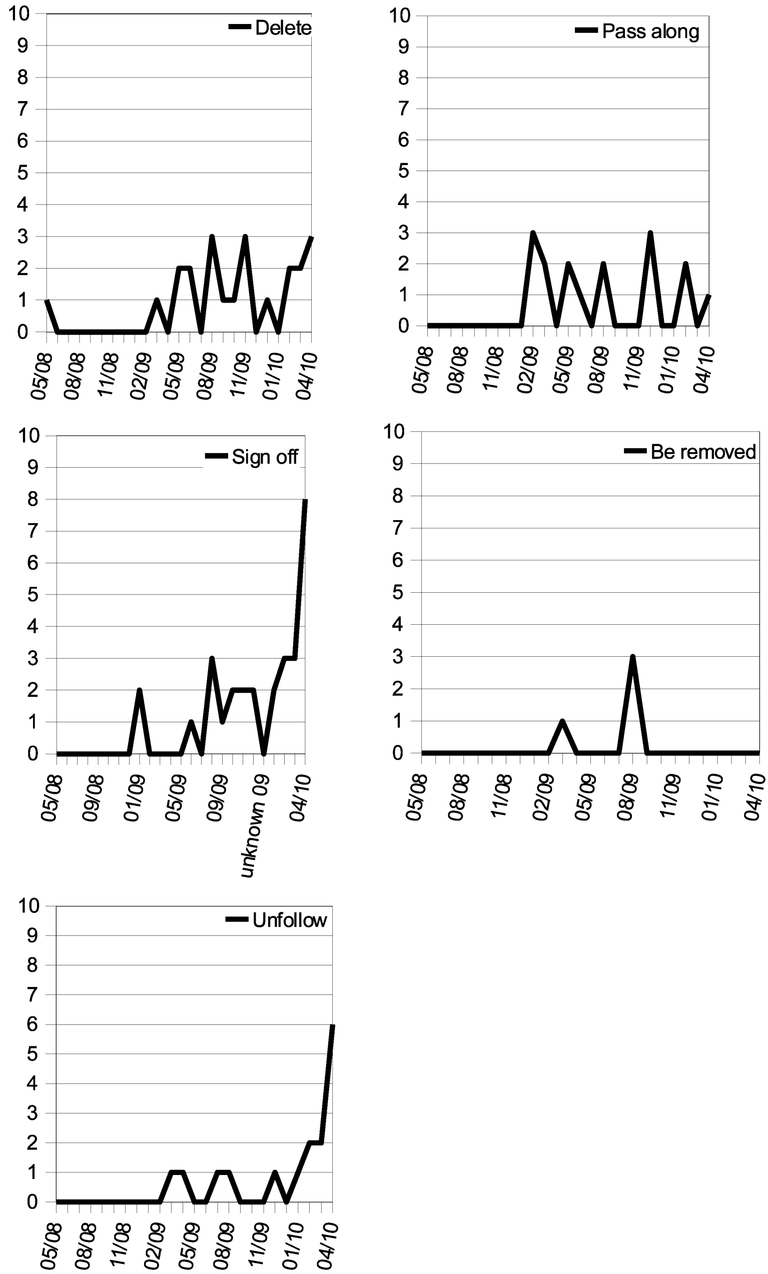

Figure 6. Monthly page frequency per assigned meaning

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 88)

As mentioned above, the rather irregular peak in August 2009 is caused by an increased frequency of *detweet* with the meaning 'being removed from Twitter'. The graph shows that except for this peak, this meaning of *detweet* has apparently not caught on and disappeared from use. The same pattern is found for 'to pass along with disapproval'. After its deliberate coinage in February 2009, an effort was made by the author to facilitate the spread of *detweet* in this particular sense. The many metalinguistic results in our data set confirm this development. However, these efforts were rather unsuccessful, since the graph shows that frequencies did not increase, but rather dropped. As Metcalf (2002: 185) notes, attempts at establishing a new word will stand a better chance if the word is 'sneaked' into the language without creating a buzz around it. Having begun its lexicalization process with the meaning of 'to delete', *detweet* has now acquired other and indeed more frequently used meanings. Its original meaning is still in use, but to a lesser degree. At the time of writing, 'to unfollow' and most notably 'to sign off' prevail. While we do not want to engage in new-word astrology, we can venture the prediction that the latter meaning will become fixed for reasons of language economy, as *unfollow* has already become conventionalized in the meaning in question, which might make a new word form for the same concept redundant.

4.4. Institutionalization

As we have mentioned, describing the diffusion of a new word in a speech community, even if it is just a limited one of the type studied here, is not just a matter of monitoring the frequency of use as discussed in Section 4.2, but also relates to the socio-pragmatic spread of a new lexical item across text-types, semantic domains and registers. Figure 7 presents a text-type analysis of occurrences of *detweet* in the five different meanings, which is based on the categories used for annotating NeoCrawler data (cf. 3.3.3).

Unsurprisingly, all meanings are used to some degree on Twitter. Specifically, 'to sign off' is frequently found in this discourse domain, because *detweeting* has become a common expression among Twitter users to indicate their upcoming off-line status. The text-type distribution, however, shows that this usage-type is by no means restricted to the microblogging genre, as *detweet* also appears in personal blogs and community portals. These three kinds of text type represent the informal end of the Internet genre continuum; other genres on the more formal side, such as news media, do not feature the word *detweet* so far, with the exception of the

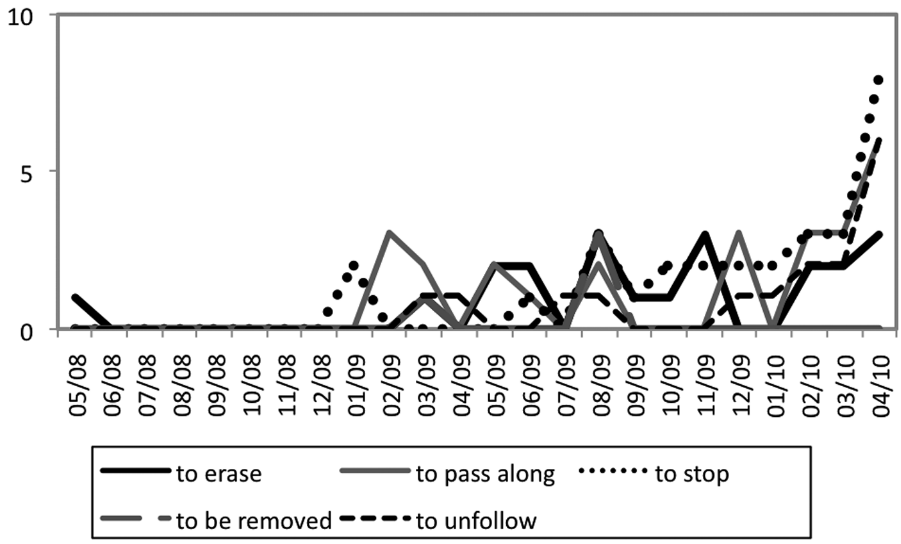

Figure 7. Overall text type distribution of detweet

Denver Post mention. This suggests that so far *detweet* has only been institutionalized somewhat tentatively, because it has not started to disperse into more formal registers and text types. It also remains doubtful whether this spread will take place at all, since the concept is, at this stage at least, used exclusively with respect to Twitter activities. It is not unlikely that its morphological make-up, i.e. the Twitter-specific base *tweet*, will prevent a future cross-over into other registers and discourse types, because of its strong cognitive association with Twitter. The current results would support this claim, but further monitoring is necessary.

Characteristic of neologisms, furthermore, is the presence of metalinguistic activity. Nearly all of the observed meanings of *detweet* have been written about and commented upon linguistically by users. Two developments can be distinguished here. In the first a metalinguistic comment is the earliest occurrence and the word is subsequently used in an object-linguistic manner. This is the case for the oldest meaning ‘to delete’. In the complementary type, the word is first used in the speech community and then commented upon at a later stage. *Detweet* with the meaning of ‘signing off’ represents this case. One of the earlier occurrences was on Twitter in June 2009. In the subsequent months, detweeting stayed under the radar of linguistic observation and did not receive metalinguistic attention until March 2010. Interestingly, it is precisely this unobtrusive, unremarked use that prevails. Although cognitively more prominent in its sense as the lexical opposite of *retweet* and actively propagated by the inventor, the meaning of 'pass along with disapproval' has not become established. Whether the presence or absence of metalinguistic comments are mere coincidental factors, or whether a significant influence on the conventionalization process exists, also constitute topics for further research.

4.5. Lexical network formation

The NeoCrawler not only allows us to investigate the diffusion of a neologism throughout the language community, registers and genres, but also to describe the lexical networks it starts to develop after its introduction. Arguably, this is an important indicator for the establishment of new words, not just from a language-systemic point of view, but also from a cognitive one, since network-building is a crucial step in lexical acquisition and the life-long reorganization of the mental lexicon (Aitchison 2003: 189–199). The following section will discuss some of the paradigmatic and syntagmatic patterns that *detweet* has already established in its early stages, which are also seen as initial evidence of the emergence of cognitive routines in the minds of language users.

In almost 30% (7 out of 22 tokens) of its occurrences with the meaning 'to delete', *detweet* is complemented by the noun *tweet*, which is of course identical in form to the base of the prefixed verb. These occurrences are all metalinguistic uses providing definitions. For *detweeted*, too, *tweet* was found to collocate in almost half of the subsequent co-texts. These observations will hardly come as a surprise, since it is only reasonable to explain the meaning of a prefixed verb with reference to its base. On the other hand, neglecting the metalinguistic function of these uses, *to detweet a tweet* can be regarded as an incipient lexical collocation or a 'cognate' verb-object construction acquiring the status of a collostruction (cf. Stefanowitsch and Gries 2003).

The restricted, metalinguistic use of *detweet* in the sense 'to pass along with disapproval' is also confirmed by the collocational analysis. Firstly, it is mainly preceded by *introducing*, which is part of the title of the article in which its coinage is explained. Secondly, the antonym *retweet* is also found in the immediate co-text, which indicates that the writer consciously tries to establish a lexical and cognitive reference to a word that is supposedly known to the readers.

The synonyms *unfollow* and *not follow* and the antonym *follow* occur in 30% of the neighbouring co-texts of *detweet* as 'to unfollow'. Collostructional preference for an object or a subject was not observed. The tokens furthermore occurred in object-linguistic use, so that the synonyms and the opposite serve as valuable cognitive and lexical anchoring points in the meaning negotiation process required by the reader.

These preliminary results indicate that since its inception, users of *detweet* have relied on strong morphological, lexical and semantic connections to the co-text. Whether and how long these initial semantico-lexical relationships are retained during the lexicalization and institutionalization process, when the need for co-textual clues is reduced due to the strengthening and disambiguation of meaning, and, more importantly, the extent of their positive or negative effect on diffusion constitute further interesting questions for future research.

5. Summary and Outlook

In this paper we have described a new methodology for the identification, retrieval and linguistic analysis of neologisms. We hope that the case study presented in Section 4 has provided an idea of the potential of the NeoCrawler for supplying the means to address long-standing questions in historical semantics and lexicology. Specifically, the case study on the neologism *detweet* has demonstrated how the NeoCrawler can facilitate the study of processes such as

– semantic disambiguation, competition-resolution and semantic change (i.e. lexicalization processes);
– semantic-grammatical correlations between word classes and meanings;
– diffusion, i.e. changes in discourse frequency;
– institutionalization, i.e. spread across text-type, genres, fields of discourse, functions (including meta-linguistic vs. object-linguistic uses);
– incipient network-formation manifested in evidence for a gradual establishing of paradigmatic and syntagmatic relations

In short, possible applications of the NeoCrawler pertain to the fields of semantic change, early morphological and grammatical change, the establishment of collocations, collostructions and valency patterns, as well as use-related aspects.

In the future, the NeoCrawler is to be optimized in a number of directions including automatic classification of fields of discourse, addition of another module to search microblogging services and extension to other search engines. Our impression is that the combination of the Discoverer and the Observer as well as reliance on the relational database approach have proven quite rewarding and promising.

References

Aitchison, Jean
2003 *Words in the Mind: An Introduction to the Mental Lexicon*, 3rd ed. Malden, MA: Blackwell.

Andrés, Louis, David Cuberes, Mame Astou Diouf and Tomás Serebrisky
2007 Diffusion of the Internet: A Cross-Country Analysis. World Bank Policy Research Paper WPS4420.

Beaugrande, Robert-Alain and Wolfgang Dressler
1981 *Introduction to Text Linguistics.* London: Longman.

Bergh, Gunnar
2005 Min (D) Ing English language data on the Web. What can Google tell us? *ICAME Journal* 29: 25–46.

Biber, Douglas
1988 *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas
1989 A typology of English texts. *Linguistics* 27: 3–43.

Biber, Douglas
1995 *Dimensions of Register Variation.* Cambridge: Cambridge University Press.

Biber, Douglas
2007 Towards a taxonomy of web registers and text types: a multidimensional analysis. In: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 109–131. Amsterdam: Rodopi.

Buchstaller, Isabelle, John R. Rickford, Elizabeth Closs Traugott, Thomas Wasow and Arnold Zwicky.
2010 The sociolinguistics of a short-lived innovation: tracing the development of quotative *all* across spoken and internet newsgroup data, *Language Variation and Change* 22: 191–219.

Carletta, Jean, Stefan Evert, Ulrich Heid, and J Kilgour
2005 The NITE XML toolkit: Data model and query language. *Language Resources and Evaluation* 39 (4): 313–334.

de Kunder, Maurice
2007 *Geschatte Grootte van het Geïndexeerde World Wide Web*. MA thesis, Tilburg University. www.dekunder.nl (accessed December 21, 2010).

(V9 8/9/11 18:34) WDG (155mm×230mm) TimesNRMT 1317 Allan pp. 59–96 1317 Allan_04_Kerremans (p. 93)

Dipper, Stefanie
2005 XML-based stand-off representation and exploitation of multi-level linguistic annotation. *Proceedings of Berliner XML Tage 2005 (BXML 2005).*

Eckart, Richard
2008 Choosing an XML database for linguistically annotated corpora. *Sprache und Datenverarbeitung* 32 (1): 7–22.

Evert, Stefan
2010 Google Web 1T 5-Grams made easy (but not for the computer). *Sixth Web as Corpus workshop (WAC-6).*

Fairon, Cédrick, Kévin Macé and Hubert Naets
2008 GlossaNet 2: a linguistic search engine for RSS-based corpora. In: *Proceedings of LREC 2008, workshop Web As Corpus (WAC4)*, Marrakesh. http://cental.fltr.ucl.ac.be/team/~ced/papers/2008-wac4-glossanet.pdf (accessed May 27, 2010).

Ferrara, Kathleen, Hans Brunner and Greg Whittemore
1991 Interactive written discourse as an emergent register. *Written Communication* 8 (1): 8–34.

Fletcher, William
2001 Concordancing the Web with KWiCFinder. *American Association for Applied Corpus Linguistics. Third North American Symposium on Corpus Linguistics and Language Teaching.* Boston, MA, 23–25 March 2001. http://kwicfinder.com/FletcherCLLT2001.pdf (accessed May 27, 2010).

Fletcher, William
2007 Concordancing the web: promise and problems, tools and techniques. In: Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 25–45. Amsterdam: Rodopi.

Ghodke, Sumukh, and Steven Bird
2008 Querying linguistic annotations. *Proceedings of the Thirteenth Australasian Document Computing Symposium.*

Hayashi, Larry, and John Hatton
2001 Combining UML, XML and relational database technologies. The best of all worlds for robust linguistic databases. *Proceedings of the IRCS Workshop on Linguistic Databases.*

Herring, Susan C.
2007 A faceted classification scheme for Computer-Mediated Discourse. *Language@internet* 4. http://www.languageatinternet.de/articles/2007/761/ (accessed June 2, 2010).

Hohenhaus, Peter
1996 *Ad-hoc-Wortbildung. Terminologie, Typologie und Theorie kreativer Wortbildung im Englischen*. Frankfurt: Peter Lang.

Hohenhaus, Peter
2006 Bouncebackability. A web-as-corpus-based study of a new formation, its interpretation, generalization/spread and subsequent decline. *SKASE Journal of Theoretical Linguistics* 3: 17–27.

Hymes, Dell
1974 *Foundations in Sociolinguistics: An Ethnographic Approach.* Philadelphia: University of Pennsylvania Press.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary
2002 XCES: An XML-based encoding standard for linguistic corpora. *LREC 2000 2nd International Conference on Language Resources and Evaluation*, Athens.

Kilgarriff, Adam
2003 Linguistic Search Engine. *Proceedings of the Shallow Processing of Large Corpora Workshop (SProLaC 2003), Corpus Linguistics 2003*. Lancaster University. http://www.kilgarriff.co.uk/Publications/2003-K-LSEsprolac.pdf (accessed May 27, 2010).

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicolas Smith
2009 *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lewandowski, Dirk
2008a A three-year study on the freshness of web search engine databases. *Journal of Information Science* 34 (6): 817–831.

Lewandowski, Dirk
2008b The retrieval effectiveness of web search engines considering results descriptions. *Journal of Documentation* 64 (6): 915–937.

Lipka, Leonhard
2002 *English Lexicology: Lexical Structure, Word Semantics and Word-formation*. Tübingen: Narr.

Luhn, Hans Peter
1958 The automatic creation of literature abstracts. *IBM Journal of Research Development* 2 (2): 159–165.

Mair, Christian
2006 *Twentieth-Century English: History, Variation and Standardization.* Cambridge: Cambridge University Press.

Metcalf, Allan
2002 *Predicting New Words*. Boston: Houghton Mifflin Company.

Murray, Denise E.
1990 CmC. *English Today* 23: 42–46.

Ntoulas, Alexandros, Junghoo Cho and Christopher Olson
2004 What's new on the Web? The evolution of the Web from a search engine perspective. http://www.cs.cmu.edu/~olston/publications/webstudy.pdf (accessed May 12, 2010).

Odlyzko, Andrew
2003 Internet growth: Myth and reality, use and abuse. *SPIE—Optical Transmission Systems and Equipment WDM Networking II*, Vol. 5247: 1–15.

Renouf, Antoinette, Andrew Kehoe and Jay Banerjee
2005 The WebCorp Search Engine. A holistic approach to web text search. *Electronic proceedings of CL2005*, University of Birmingham. http://www.webcorp.org.uk/publications.html (accessed May 27, 2010).

Schmid, Hans-Jörg
2011 *English Morphology and Word-formation: An Introduction*, 2nd ed. Berlin: Erich Schmidt Verlag.

Štekauer, Pavol
2002 On the theory of neologisms and nonce-formations. *Australian Journal of Linguistics* 22 (1): 97–112.

Tournier, Jean
1985 *Introduction Descriptive à la Lexicogénétique de l'Anglais Contemporain*. Paris: Champion-Slatkine.

Uyar, A.
2009 Investigation of the accuracy of search engine hit counts. Journal of Information Science 35 (4): 469–480.

Werlich, Egon
1976 *A Text Grammar of English.* Heidelberg: Quelle and Meyer.