INVESTIGATIONES LINGUISTICAE, VOL. XVI; POZNAŃ, DECEMBER 2008

Comparison of selected methods for the retrieval of neologisms

Piotr Paryzek

Institute of Linguistics, Adam Mickiewicz University
al. Niepodległości 4, 61-874 Poznań, Poland

piotr@pparyzek.pl

Abstract

The paper discusses and compares several semi-automatic methods used to extract neologisms from linguistic corpora. All the methods are based on the concept of discriminants, or textual features (both lexis and punctuation), that either precede (lexical discriminants) or confine (punctuation discriminants) phrases in which the occurrence of neologisms is higher than elsewhere in the text. Excerption and comparison was conducted on a corpus of 45 million words, articles from *Nature* scientific magazine. The putative neologisms were extracted using morphological analysis and frequency of their occurrence in the Google search engine. The result is a list of 1000 neologisms and assessment of the efficacy of each method.

1 Introduction

Now, with the ever increasing pace of progress in science and technology, there is a vast number of neologisms that should be recorded by lexicography. The coinage of neologisms is especially evident in English, hence the scope of this paper.

Excerption of neologisms[1] has traditionally[2] been effected manually. A skilled person, possibly a lexicographer, reads a text and puts down unknown lexical units which are subsequently verified against selected dictionaries, their spelling and relevance (a fairly subjective stage) is checked and a list of neologisms is produced. Albeit potentially highly accurate[3] (all new lexical units in a text or collection of tests are excerpted), the method is uneconomical, for the volume of published texts prevents their thorough analysis or any analysis whatsoever.

Therefore, there is a need of automation in excerption studies, i.e. methods that will facilitate the process, reduce its duration and limit the need of human contribution.

Automation in lexicography, introduced by engineers rather than lexicographers, has been especially successful in the area of collocations and spelling correction (Dias 2000, Golding, Schabes 1996, Gries, Stefanowitsch 2004). Current methods[4] employ highly sophisticated mathematical apparatus to produce lists of collocations[5] that should later be examined by lexicographers as to their relevance, which, unfortunately, is rarely the case.

[1] Cf. Buttler 1962, 1993, Wawrzyńczyk 1994, 1999, Smółkowa 2001, Stoberski 1976.
[2] Cf. *Wstęp* w *Słowniku języka polskiego* ed. by W. Doroszewski (1958-1969), cf. Bańko 2001.
[3] Cf. results of manual excerption in: Wawrzyńczyk 2000.
[4] Cf. Siepmann 2005.
[5] Cf. Buczyński 2004, Moszczyński 2005

With respect to the excerption of neologisms, however, statistical calculations are less promising, for neologisms are single and unique lexical units (i.e. individual words – *hapax legomenon*) and any data on their frequency will not be much informative, because an interesting (from a lexicographic perspective) neologism may well occur only once in a text or collection of texts (corpus) and its existence and location cannot be predicted using mathematical methods.

Enter linguistics, as a replacement of the statistical approach. The concept was formulated in Chlebda (1991) who noted that phrasemes[6] occur within quotation marks or after certain phrases (such as: *tak zwany, jak to się mówi* in Polish). This idea has been expanded by P. Wierzchoń and translated into a automated or semi-automated method for the excerption of neologisms. The method centers on the neighborhood of a lexical item marked as neologism. Thus, according to the first hypothesis, neologisms are found (at least with greater frequency) within quotation marks, i.e. phrases confined by quotation marks are treated as the input for subsequent analyses (the method has been treated in detail in Wierzchoń 2003, 2005a). The other concept is based on what usually precedes neologisms, that is language-specific phrases, such as *so-called* etc. Theoretical background and practical instructions on how to retrieve these phrases (with particular emphasis on inflecting languages[7], such as Polish, which call for the extensive use of regular expressions) have been presented in Wierzchoń 2002.

Both types of methods take advantage of morphological analysis8, which itself can be an independent method for the excerption of neologisms.

The aim of this paper is to test the methods on the scientific register of English in the form of texts published within the past several years (expected to contain a large number of neologisms), compare them (and provide statistical data on their performance) and automatically reduce the number of words in the final list by checking the frequency of their occurrence on the Internet using a search engine.

2 Premises

2.1 Corpus

The corpus[9], or collection of texts, used as input in the proposed method should be selected so as to maximize the number of excerpted neologisms, defined as units most probably not yet registered by lexicographers and surely not registered in dictionaries, such as the *Oxford English Dictionary* or *Webster's Third New International Dictionary*. Thus, two linguistic perspectives emerge: diachronic and synchronic, or old and latest texts, for the occurrence of forms new to lexicography (if archaic in terms of their place in modern linguistic systems) is to be registered and analyzed within the method. Of the two, the latter approach is definitely more interesting for lexicographic and general linguistic purposes as it enables observation of the latest developments in language and registration of new forms almost *in statu nascendi*[10].

[6] Cf. the definitino of *phraseme*: "every linguistic sign, irrespective of its semantic status and formal structure, which constitutes a name for a content potential (term) referred (uttered) by a speaker as its relatively constant symbol." (Chlebda 1991: 27).

[7] Inflecting languages, i.e. those using inflections, or affixes (morphemes that denote specific grammatical relations, such as gender, number, case etc.; one affix often denotes more than one grammatical category).

[8] Morphological analysis enables automatic rejection of forms known to the analyzer and thus not interesting for excerption purposes.

[9] An intuitive definition of the corpus is adopted at this point, i.e. a collection of texts.

[10] Furthermore, morphological analysis of old texts will likely be problematic (too many words marked as unknown in the absence of an analyzer intended for such texts).

In consequence, texts in the corpus should meet both of the following conditions:

1) latest texts (i.e. those published recently)[11];
2) texts whereby a lot of neologisms are likely to be used; one of the most promising text types are scientific journals[12], especially those treating of natural science, i.e. those disciplines where progress is most noticeable.

2.2 Retrieval of phrases

The fundamental question is: Is it possible to find graphical (non-lexical), lexical or other specific, easily retrievable entities that precede or contain phrases where neologisms occur with greater frequency than elsewhere in the text? Is it possible to automate the process? The method proposed in the article is centered on the concept of **discriminants** in the neighborhood of which neologisms tend to occur. For the most part their existence is language-independent, i.e. even if their graphical forms, or vocabulary used, are different, they should in principle exist in any language, which ensures that the method is expandable to many other languages.

There are two types of discriminants: lexical and punctuation ones. Lexical discriminants are phrases that usually, or with greater likelihood, precede neologisms. In a way, they announce neologisms or define the new or unknown.

As the working language for the project is English, the following phrases have been selected to verify whether they are indeed advantageous compared to the random sample:

- *termed*,
- *called* (which includes *so*(-)*called*),
- *known as*,
- *defined as*.

The other type is punctuation discriminants. They are expected to confine phrases likely to contain neologisms. Two candidate types of punctuation marks have been selected: single and double quotes. It will be verified whether they indeed announce neologisms and whether either discriminant is better (if so, it would be possible to demonstrate the differences and regularities in their usage).

2.3 Morphological analysis

Morphological analysis is a procedure that lemmatizes a form of a word as it appears in a text and assigns to the stem symbols that describe the form. This is accomplished by computer programs called *morphological analyzers*[13]. In this study we are not interested in what specific form has been used in the text, but rather what forms are unknown to the analyzer[14]. While all neologisms should fall into this group, not all elements unknown to the morphological analyzer will be interesting for lexicographic purposes. For the analyzer's capabilities are limited and many commonly used words may be marked as unknown.

2.4 Extraction of the rarest units

Wordlists generated during morphological analysis depend solely on the morphological analyzer used. Thus, many of the words marked as unknown are likely not to be interesting for lexicographic purposes; even though marked as such, they could hardly be called neologisms[15]. Therefore, there is a need to find an automatic tool that would select the pool

[11] So as to find new constructs not yet registered.
[12] Apart from literary texts written by authors who tend to employ neologisms. These coinages, however, while interesting, may not be incorporated to the language.
[13] Cf. Bień, Szafran 2001.
[14] The AOT analyzer, available at www.aot.ru, has been used in the present work.
[15] Such as *baseline, assistive, apoptosis, adversely, luminance, decreasingly, cutoff, adenosine, artificially, angiogenesis*.

of the most interesting units. At best, the resulting list would provide a ranking of words. How to rank words? One of the most reliable methods (apart from those based on balanced corpora, such as the British National Corpus) is to estimate the frequency of a word using the Internet, or a web indexer whose most useful feature is that it lists the number of occurrences of any word in the indexed webpages.

2.5 *Random* file

In order to establish absolute efficiency of the method (and verify whether it is advantageous compared to the random selection of words for excerption purposes) and respective discriminants, a *random* file has been created, a sample of the whole text. It serves as a reference to which other files are compared.

3 Method

3.1 The data

The data used in the analysis is a collection of complete articles including *Editorials*, *Research Highlights*, *News*, *News Features*, *Correspondence*, *News and Views*, *Brief Communications*, *Articles* and *Letters*, published in scientific journal *Nature* between 1997 and 2005 (a total of nearly 450 issues) and available from their website as html files. Table 1 below presents details of respective years' issues.

The journal has been selected for the following reasons: 1) it presents the state of the art in science with particular emphasis on fast-growing disciplines, such as biological sciences, in which a significant number of new concepts and entities requires plenty of neologisms to be coined; 2) being one of the most respected scientific journals (with one of the highest impact factors), it publishes peer-reviewed articles, which is likely to contribute to their factual and linguistic correctness (also with respect to neologism formation); 3) the journal is published on a weekly basis, i.e. should contribute a large body of data; 4) the data are available in the html format, easily convertible to the txt format used in further processing.

Table 1. The characteristics of the corpus

Year	File size, MB[16]	Number of words, in thousands[17]
1997[18]	17.89	2737
1998	30.59	4703
1999	29.47	4511
2000	35.28	5412
2001	37.40	5736
2002	33.00	5042
2003	32.57	4985
2004	36.56	5611
2005	39.39	6003
Total	292.15	44740

[16] In text files.
[17] In text files.
[18] The significantly smaller volume of 1997 year's issue is due to the fact that a number of issues were not available in the html format and thus were not included in the corpus.

3.2 Data preparation

The following operations were conducted to prepare the data:

1. Merging the html files (one html file contains one text) into a large file.
2. Clean-up of files to remove all the recurring elements, i.e. other than articles proper, such as copyright notices and links not within the article body. Bibliography has been left unchanged.
3. Removing html tags.
4. Converting the large html files (individual year's issues) into pure non-html txt files (9 files total).

3.3 Retrieval of phrases

In the case of putative lexical discriminants of neologisms (*termed*, *called*, *defined a*s and *known as*) retrieved were phrases starting with a discriminant and ending with a proximal period.

In the case of punctuation discriminants (single and double quotes), the whole phrase within the quotes was retrieved. The processing of data was conducted separately on each year's issue, which resulted in 55 files total (6 discriminants multiplied by 9 years + *random* file).

Spaces were inserted between words and reference marks (i.e. between the last letter of a word and the first number in all phrases), such as in the following sentence:
Slow group velocity was measured recently in photonic crystal structures with ultrafast pulse propagation techniques15, 16.

This introduced spaces into units, such as H5N1 (H 5 N 1 after the operation), but as the scope of the analysis is limited to lexical units with numbers being of no importance, this step can't have distorted the data.

The retrieval operations were in most cases straightforward (carried out using simple syntax of regular expressions), with the notable exception of single quotes which are difficult to distinguish from apostrophes (especially those marking plural Saxon genitives).

The *random* file has been created as follows:

- based on the 2005 year's issue;
- all blank lines were removed;
- resulting number of lines: 94,000;
- every 1250th line selected for further analysis[19].

This step produced the following results (example, 2005 year's issue):

1. termed:

termed extrinsic noise), and the results suggested that some components of extrinsic noise affect gene expression in general 11.

termed duplication shadowing, suggests that loci near clusters of segmental duplication may be more susceptible to duplication deletion, probably due to an increased frequency of non allelic homologous recombination 1 8.

termed lipoproteins are a major constituent of these extracellular compartments 3, yet their role in CD 1 antigen presentation has not been investigated.

termed slow dynamics 23, meaning that the modulus slowly returns to equilibrium over several hours or even days after the wave energy has disappeared.

[19] Thus, an easily analyzable sample representing the whole text has been prepared (5461 words) and further processed like the other files.

termed the BHC or BRAF HDAC complex, which is required for the repression of neuronal specific genes 1.

2. Known as:
known as the particle.
known as ACE D, makes more ACE than the other common version, ACE I.
known as myogenesis, begins in transient blocks of tissue called somites.
known as British India may be a country for political purposes, but in no proper sense of the word do they constitute a nation.
known as brownian motion – also heralded a revolution in physical thought.

3. Defined as:
defined as being connected if any of their constituent nodes are linked.
defined as 2 metres or more in length) were discovered, starting in 1828 and ending in 1996.
defined as that of the normal yellow gene in D.
defined as tannins, also precipitate proteins and are perceived as astringent.
defined as being due to relatively stable changes in gene expression without changes in the DNA sequence of the gene.

4. Called:
called the Cubiculum of the Ocean.
called siderophores, which are used by bacteria to absorb iron, a nutrient they need to produce essential enzymes.
called AHLs, or N-acylhomoserine lactones.
called circumstellar disks, and to look at disks ranging from a million years old up to the age of the Sun is to look at the planetary construction process.
called the triple process.

5. Single quotes:
'instruments'
'Lisbon objectives'
'*junior Nobel prize*'
'*biospherians*'
'*Sistine Chapel of its time*'

6. Double quotes:
"*With world attention focused on natural disasters, it is an idea that many people feel is ripe,*"
"*This needs to be put together now,*"
"*There are a large number of bodies already doing this. We need to pull things together under a single umbrella,*"
"*When scientific bodies tell governments what to do they get rejected,*"
"*Unless there is a supplemental appropriation, then the dollars pledged will definitely have to come out of current budgets and thus will compete with other needs,*"

Manual analysis of these phrases reveals certain interesting units, such as *biospherians* (single quotes), *somites* (known as) and *acylhomoserine*.

Table 2 below presents the results of phrase retrieval (total for all year's issues).

Table 2. Summary of the retrieval of phrases

Discriminant	Number of occurrences	Number of words total	Avg. number of words per occurrence
termed	1014	13685	13.50
called	8067	111977	13.88
defined as	1356	24649	18.18
known as	4178	59796	14.31
single quotes	53949	95506	1.77
double quotes	45457	610412	13.43

Double quotes provide the greatest volume of data and single quotes the greatest number of occurrences and the shortest phrases, which will prove very efficient at later stages.

3.4 Morphological analysis

One of two steps intended to excerpt neologisms from files acquired during retrieval of phrases, morphological analysis is aimed to isolate lexemes not recognized by the morphological analyzer (i.e. marked as unknown). This should significantly reduce the number of words and provide input for the subsequent step.

The files, after necessary adjustments (e.g. slashes, apostrophes and hyphens were converted to spaces)[20] were processed by the morphological analyzer. Based on the resulting files, lists of words not recognized by the analyzer were derived.

The lists were further processed:

1. All words containing at least one capital letter were discarded to remove names, proper names, acronyms etc. At the same time all potential neologisms written with a capital letter were lost. However, as the aim was to automate analysis, this operation seems to be justified.
2. All words containing one or two letters were discarded, such as single *s*'s originating from the replacement of apostrophes by spaces. This cannot have had any adverse influence on the excerption of neologisms.
3. The resulting word lists were ordered alphabetically and repetitions were removed.

Listed below are examples of results produced in this step (2005 year's issue):

termed:

acetoxonium, adenomatous, allelic, allodynia, anammox, antagomirs, antigenic, arteriogenesis, autoimmunity, biodiversity, blastospores, calmodulin, cannabinoid, catenin, chemicurrent, conpats, copaxone, cyclin, deimination, doxycycline, ecogenomics, enteropneusts, epistasis, et, eukaryotic, exchanger, extracellular, fru, genomic, genomics, glatiramer, haplotype, hemichordates, hemifusion, idiomorphs, inducible, interswitch, kinase, lipoproteins, lophophore, lymphoid, magnetoelectrics, megakaryocytes, metabolator, metagenomics, micromanipulation, minisleep, mitochondrial, mns, molecularly…

called:

acetylcholinesterases, acron, acrosome, acylhomoserine, affinities, aminoglycosides, amphipaths, anaphylatoxin, aneuploidy, angiogenesis, angiotensin, anomalously,

[20] The morphological analyzer treats units containing slashes or hyphens (such as and/or or much-disputed) as words and often marks them as unknown (such as *risk-reduction* (*schemes*) or *earthquake-prone* (*areas*)).

antiporters, apolipoprotein, appressorium, aquaporin, arbuscular, aren, argosomes, artesunate, astrocytes, autoimmunity, autoionization, barcoding, bedforms, benztropine, bevacizumab, biotech, biseparable, blastocyst, blastocysts, boreoeutherian, brainstem, branes, bromodomain, businesses, calmodulin, cardiospheres, cathodoluminescence, ceftazidime, cephala, chamosite…

defined as:
blastocoel, cutoff, cyclin, decreasingly, decribed, defensin, desmethyl, distally, euthanization, fluence, hindcasts, hopane, ischaemic, kilobase, luminance, mainshock, mers, min, myocyte, nanotech, non, normally, orthologue, pixel, positionstart, positionsteady, postsynaptic, pre, predominantly, qualitatively,, seroconversion, shuttings, tastant, tetramer, timestart, timesteady, transcriptional, utc, viraemia.

known as:
achiral, adversely, allorecognition, anammox, androdioecy, angiogenesis, anthracotheres, antiepileptic, antigenic, apoptosis, apoptotic, arbuscular, archaea, arguably, arogyapacha, aspergillus, assistive, astrocyte, backarc, betalains, bevacizumab, biodiversity, biomolecular, biosynthetically, blastocyst, bonannione, bonobo, brevetoxins, brevis, buffelgrass, cardioprotective…

single quotes:
abcd, acetoxonium, achiral, acron, actin, adakite, adaptationism, addback, adenosine, adipocrine, adipokines, aflatoxin, afterglows, aggrecanase, aggrecanases, aldol, allostatic, analyte, anammox,, angiogenesis, angiogenic, anomalously, antagomirs, anthropodenial, antibias, antigenic, apo, apoptosis, apoptotic, apsidal, archived, artisanal, aubotsy, autoantigens, autotoxicus, barcodes, baseline, bedform, bedrest, biconical…

double quotes:
abstr, abundantly, accretionary, actin, adenosine, administratively, aediculatus, afferents, airsurfers, albicans, alloimmunity, aminobutyric, amplicons, amu, anarcho, ands, angiogenesis, angiogenic, angiopoietin, antagomirs, anticancer, anticontributive, antiferromagnet, antivivisectionists, antonin, aoml, apoptosis, apoptotic, apos, aquaculture, archaeal, archaeon, archivable, arcsec, artesunate, artificially, arxiv, astrocyte, astrocytes, astrocytogenesis, astrometric, atonia, auflosung, autosomal, axonal, bacterioplankton, bandgap, barcode, barcoding, barite, baroclinic…

The results of this step are presented in table 3 (total for all year's issues).

Table 3. Results of morphological analysis

Discriminant	Number of unique words marked as unrecognized
termed	528
called	2614
defined as	364
known as	1614
single quotes	3915
double quotes	3530
Total	**12565**

As we can see, the number of words has been greatly reduced with respect to the preceding step, especially in the case of double quotes.

3.5 Isolation of the rarest units

As our aim is to isolate not only neologisms as such, but at best units not yet registered in lexicons, Internet resources have been used to provide estimates as to what units are most likely new to the language and lexicography. The gauge has been the rarity of a word. To that end, the Google search engine was used to assign the number of occurrences of a given unit in the websites indexed by Google in the following format: [no. of occurrences] [unit]. Subsequently, the list was sorted numerically[21].

As for the threshold number of occurrences for an entity to be considered interesting as a possible neologism, the value of 1000 occurrences has been adopted. This makes it possible to severely limit the number of units, but at the same time select only the rarest. While the value is purely arbitrary, the resulting entities contribute a set being extremely interesting from the lexicographic perspective, also with respect to word-formation patterns. The results of this step haven't been presented in a table; see the final results after the subsequent step.

3.6 Manual analysis of word lists

Even though all the preceding steps were highly automated, the final analysis of word lists has to be conducted by a linguist or a native speaker to pinpoint all the cases of typographic errors, foreign words and ephemeral[22] forms. Listed below are examples of units that have been regarded as unsuitable and thus removed from the final lists:

- typographic errors: *theoreticalpractice, expertiseof, dignityof, appealedto, humandignity, believesit, organizedaround, whichgovern, marginalenvironmental, alwaysalso, preposterousconclusions, beethically, connectivetissue, findingsis, directedtothe, supranationalinstitute, priority, difficult, ofcompounds, strontium,*
- explanation of patterns existing in other languages: *mouseeats, mousegoes;*
- symbols: *uvcalc, fstim, sqtz, ndisl,*
- foreign words: *shuvuu, sötted, génopôles, entwicklungsbiologischer, betsika, augebitur, pertransibunt, yanzigou, zerstückte, subitaneis, iigiracóobitooree, iigiracóobiwareec, sorokinii, maúudabi, semicelatum, biostratigraphische, maagaríshdawacee, mosbachensis, carnegii, wiáha, macée,*
- archaic forms: *burlesqt, heareing, equalle,*
- ephemeral forms: kvestion, vhatever, physicists, discuzzed, failurez.

4 Final result

The words listed below are the final result of the proposed method divided into respective discriminants and year's issues.

1997

called

trochleated, rosettins, tagamites, mertensian, palaeoceanographers, magnetostrophic, lepidotrichia, orviétan, prosaccades, australopiths;

[21] The analysis using Google was conducted in March 2006. Web indexing is an ongoing process; therefore, the exact number of occurrences varies with date (and also with location).

[22] Ephemeral, i.e. forms used only once, e.g. to reflect incorrect pronunciation or slip of the tongue.

double quotes
prowbley, naturify, anandamidergic, polyphyrin, commaform, superpenumbral,, uneconomy, recorrecting, cathechins, homeodynamic, pseudomedical, thatled, megabillion, radioastronomers, presqualene, governmentalist;

single quotes
holotheres, synstorm, klinorhynchy, arthropodization, superdimer, oligomolecular, mertensian, eupantotheres, sigmage, supervolatile, megapore, paranotal, parareptiles, neurophilosophers, autapse, transducisome, copulators, superministry, medusoids, balloncraft, framboid, connectoplasm, bonebeds, micromoulding, monophagy, nanofossils;

defined as
vaverage;

known as
asioryctitheres, zalambdalestid, adenotin, epicathechin, lucibufagins, epigallocathechin, cathechins, epipubic;

termed
presqualene;

1998

called
cyproase, bifurcationists, aplanktonic, parapsid, hexabrachions, waiverers, haemoglobinase, osteolepiforms, micropolygyria, rhipidistians, lexitropsins, polyamorphism, aseismically;
double quotes
palaeopenetrometers, axoniform, dehomologation, diskoseismic, biotolerance, megamullions, megamullion, systematicists, spiritdom, outreproduce, preformationists, radioastronomers, chromatosome, misappliance, scapulocoracoid;

single quotes
macrosyllable, piezonail, altoradiometer, ennobelled, diplosyllables, lepiform, platigem, preprismatic, tachopause, mutasomal, pseudodating, bifurcationists, sapromyiophily, aplanktonic, necrolab, cuspier, parapsid, polynail, precompensates, cephalobid, sphingophily, neontologist, cavicapture, neoglycopolymers, megabeds, gastrophysics, pseudodate, regularist, megaturbidite, conciliance, megaturbidites, lettersound, synneusis, baryometer, osteolepiforms, tetherable, lepospondyl, palaeothermometer, rhamphorhynchoid, osteolepiform, pyracylene, vendobiont, tunnelized, superplanets, megachannels, polyamorphic, osteolepiformes, downwelled, transducisome, bakedness, pinscape, anthracosaurs, segnosaurs, actinosporean, biofluiddynamics, etchability, echeme, ornithophily, rothomagensis, intrasteric, exflagellation, kleptoparasites;

defined as

known as
loxommatids, baphetids, friarbirds;

termed
megabeds, megaturbidites, megaplumes, exflagellation;

1999

called
dygments, pseudocopulations, isochelae, antishocks, dimycocerosate, gabbronorites, protohaem, formatrix, pyrobitumen, haemangioblastomas;

double quotes
scattercirrus, fallnimbus, ascendstratus, foldedcumulus, reindigenizing, upcruitment, tribosphenid, gelbrain, mimeomorphic, photofootprint, zooblot, mesdemet, wrongedy, androgynization, ultracivilized, palaeohydrogeology, adamantoid, selfplex, subitize, decruitment, transgenomics, formatrix, volvelles, unmixedness, pasteurien;

single quotes
dygments, pongidized, palaeolandslide, pathophage, neurotrophinergic, reindigenizing, slabology, geohopanes, yellowfix, polyplocodont, microdermic, tidalists, micromastic, untransfectable, asthenoliths, hypobradytely, protolarva, regressins, chaincloth, palaeoamericans, rollertube, pellatron, pseudocopulations, petrophage, highconic, embryonization, superwells, operomics, pseudized, rhamphorhynchoids, legness, antishocks, crosspriming, comodulated, mudsplashes, segnosaurs, demultiply, superswells, anthophyte, beerstone, prosegments, inchworming, mudsplash, triconodont, cornealis, cratonization, deoxygenase, uninodal, miniwar, superalliance, gabbronorites, pseudosubstrates, polyamorphism, eupyrene, pinchase, ecdysozoan, pasteurien;

defined as
trenchward;

known as
tidalists, porolepiforms, gongylidia, aspartases, clathrasils, eutely, syntrophs, mycetocytes;

termed
methyldiamantane, prosegments, pseudolysogeny;
2000

called
nitroimidazofurans, chemotrophism, bisporphyrinate, mnemiopsin, berovin, photoresolution, mitrocomin, nitroimidazopyrans, phialidin, solardomes, ventists, haemangioblasts, palaetiological, presenilinase, fimbral, enterohaemolysin, unicolonial, magnetochiral, specillum;

double quotes
dyschromatopsics, superannalist, malaricissima, intrabranally, nanostencilled, telanthropoids, dysmenorrhaea, arttaste, astrometers, pocilloporins, pagodane, ventists, chailer, supermeme, palaeobiologists, chailing, perflation, radiocollar, foistered, hydrinos, unrenounceable, macroelectronics;

single quotes
supersupershifted, micrometozoan, staygreens, membrasome, intrabranally, physiopole, cosmuck, interdimers, antigeroid, scientaria, unfaults, postfus, electrofoil, secretasome, telonomic, lymphapophyses, amphidromies, rabbitized, uncopying, secretosome, peristriatal, abgerminal, attolitre, morphodynamically, microbiography, superconvection, cyclosynchrotron, unfoldases, misprojections, supergreenhouse, commodifaction, nanoprojects, antitestis, antimimetic, neurorobotic, garbenschiefer, genocopy, rhamphorhynchoid, solardomes, transcriptosomes, osteoimmunology, chailed, maxizyme, retrohoming, magnetochiral, paralemniscal, tripledecker, equilibriation, garnetite, centauron, nanorover, quarktet, unfoldase, syntrophy, chaostory, axonopathies;

defined as
backlabelled, interprotomer;

known as
malaricissima, lamillipodia, haemangioblasts, retrohoming, specillum, equilibriation;

termed
euagaric, thelephoroid, pseudopupil, repressilator, allospecificity, syntrophy;

2001

called
vermilarva, oomicides, archidynamic, exterlibral, coordinometer, trichromator, gonialblasts, cerebrotype, hippopotamid, ornithurines, mediatophore, ferropericlase, paranematic, mitosome, abembryonic, chargons, elaiosomes, chargon, volicitin, paramagnons;

double quotes
mathematicability, vermilarva, tribosphenidans, unrecognizedly, nitrosohydroxylamines, eupantotheres, supersog, diskoseismic, nonhumanistic, energeticists, canaliform, cichlidiots, symmetrodonts, anthropodenial, symmetrodont, phytoanticipins, phenologists, turnipy, ornithurine, unsealable, superprotonic, unakin, palaeodata, fossilists, alkaptonuric, formatrix;

single quotes
palaeosterilization, oomicides, ovasomagenesis, tribotheres, prefacilitate, ascohymenials, extracleithrum, ormiaphones, mutasomal, exosemiconductors, heteropolyblues, metastatistic, electrosomatic, baritization, nitrosoreductase, superphyletic, ovumsum, megabasins, leafpoints, transducisomes, eupantotheres, megabasin, supersog, cryptidin, cerebrotypes, piscidins, zhelestids, metasaccharinic, gliriform, guardees, quasicondensates, dealloyed, shrimpoluminescence, crownward, optipulse, crystallomics, polypodiaceous, chronotherapies, tetrahedrality, argosomes, electrolamp, orbitons, aurophilicity, endovanilloids, ovasome, rumbleometer, crosspriming, mudsplashes, signalplex, transducisome, glucolipotoxicity, mitosome, automone, superministry, polyplacophoran, microislands, abembryonic, guardee, deoxygenase, superhybrid, resublime, extremozymes, elaiosomes, □ndantino, chargons, cytonemes, calfuse, saccharinic, chargon, pinchase, slushball, preformationist, condylarths, tectosphere, immunoediting;

defined as
ultradivided, zonasulcate;

known as
homomeroquinene, tribosphenidans, asomatopagnosia, superbrownian, flexoelectricity;

termed
cataflexi, hyperflexistyled, exosemiconductors, cytonemes;

2002

called
ethesiometer, controlniks, undecouplable, complexomics, prodiginine, ladderanes, trimethylmethoxysilane, gamergates, oligopyrrole, photoheterotrophy, mitosome, nanowhisker, thermochromatography, lipochitin;

double quotes
zeppelinoids, unchemist, chemoinformatician, cerebrotype, paramilitia, autocreative, chromicized, obsessionism, microlepton, baryometer, brainspinning, aerosolizable, photoheterotrophy, anthropozoic, brainedness, anammoxidans, electrions, biofraud,

single quotes
protoanthracosaurs, muonionization, palaeophosphatometry, bradyfauna, equivalogue, phosphatometer, transloxer, pseudohaemolysin, controlniks, avnosmia, undecouplable, coelibactin, spexels, supercorrelation, hoxology, unifolds, complexomics, stromatoloids, coelichelin, printspeak, peytoia, microchimaerism, antiuricosuric, microleptons, nanothermometers, resulphurized, panchabhoota, vaccinomics, pseudodimer, baryometer, hyperscanning, distalized, lepospondyl, pseudomagnetic, dihapto, photoaptamer, osteodontokeratic, superchemistry, ladderane, divisome, anammoxidans, anthracosaurs, timesome, triconodonts, unclumped, inchworming, aftercontractions, nanothermometer, biomotors, dispersalist, megaregolith, repressilator, metacarbonate, transdifferentiating, paramorphs, slushball, tomographer;

defined as
ladderane, tailbeat;

known as
flexibilatis, somatoaxon, cornutes, mitrates, lamellipods, minifilaments, muscivorus, pericentromere;

termed
pseudohaemolysin, epiparasites, fluorophosphine;

2003

called
homodisciplinary, aplanktonic, osteochondroprogenitor, lorisiforms, missionnum, overdeepenings, phaseonium, routinizable, nucleofilaments, neurocrystalline,

microsyntenic, interglomerular, cytonemes, mitosomes, ladderane, shavenbaby, enteropneusts, strepsirrhines, spectrosome, nanoprisms, pentagonally;

double quotes

nanolecture, tracheals, retrovaccinology, exosymbiotic, eutelic, oresmen, vegetalization, supernebula, superswells, neurocrystalline, superinvar, intramers, immolative, chronophotograph, aëroplane, heliocentricism, nanolitres;

single quotes

hibbenisms, colliculoreticulospinal, palaeopasteurization, chaperonology, heterodisciplinary, neohubbertarians, nanoharvesting, glottoclock, protoconoid, nanolouse, incommensuracy, scoopophobia, poppase, aplanktonic, rostralia, tauidion, eucentricity, isoindene, toothcombed, missionnum, duails, fluorobodies, dephosphins, superwetting, functionation, polyaxonal, overdeepenings, brodae, antiglass, supersegment, baroplastics, synaptotoxic, phaseonium, routinizable, supernebula, sonocytology, posteriorized, ubiquityl, subjunctional, paraspeckles, ladderane, xenoscience, preferers, phytometer, gammation, cuckolders, biobugs, aëroplane, phluorin, plantibody, dedifferentiates, lovespot, propiece, attophysics, degradome, microacoustics, cuckolder, retrotranslocated, gerontogenes, nanopod, nanocage, toplighting, meltback, pinchase, sulphenyl, bonebeds, regassed, probablys, microstreaming, tectosphere, systemicity;

defined as

vinculinaggresomes, opistodontians, sphenodonts, doliolaria;

known as

anthracotheriid, afrotheres, ambulacraria, vegetalization, glaciohydraulic, ladderanes, hypobranchials, hyperstriatal, doliolaria, lysenin, enteropneusts, malariotherapy, ceratobranchials;

termed

polyaxonal, baroplastics, cryptospores, paraspeckles, plastochrons, immolative, cuckolder;

2004

called

polymorula, quduties, thaxtomins, paranotal, chordamesodermal, palmitoylputrescine, lasetron, polyhook, undruggable, oosome, telepreventive, phosphoroamidite, heterotachy, magnetoelectrics, geopoetry, sageing, mitosomes, destabilase, intramers, haemangioblast, superlubricity, meristemoid, hanatoxin, batrachotoxins, monophagy;

double quotes

mouthwashology, thermodramatics, gerontocractic, dodaersen, yeastification, tensegral, antivaccinationism, progressible, uninvitation, sunklands, concestor, geopoetry, coethnic, cryptofauna, multicontinental;

single quotes

innateosome, opportunitroph, nanosalt, nanocrimelab, techniquities, apternodontids, brachylabic, nanosociology, superpostdocs, merotely, macrolabic, morphogeologic, hyperlethality, minifacets, degelled, diftox, multirhythm, yeastification, alcosols,

rheoreversible, nonsolvers, metaethnic, foldhunter, camgaroos, hypolithon, hypoliths, countergradients, helixhunter, anoxicity, intervality, norhipposudoric, monoreceptor, nanaerobes, microdiverse, overdifferentiated, hipposudoric, receptorology, boxological, mussid, photocaged, megaturbidite, amphitelic, anthropodenial, calcichordates, lasetron, hyperspecificity, demethylimination, ignorosphere, osteolepiforms, palaeothermometer, telepreventive, undruggable, sphenosuchian, macroelectronic, tilepath, uninvitation, softwiring, intramers, syntelic, sideground, sageing, prehairpin, antagomirs, altriciality, glassformers, faviids, cryptofauna, arctometatarsalian, fermionized, preorganize, polyheads, superlubricity, faviid, cainism, thunniform, asabiya, ultracontigs, erectines, megamouse, nanisani, polyamorphism, exteins, deiminated, instructionist, myoseptum, syntrophy;

defined as

polygerm, hypolithon, hypoliths;

known as

cantharidiphilous, unhedgehog, cinctans, ctenocystoids, pintronic, tarsioids, homalozoans, trioxolanes, siderocalin, uterocalin, stylophorans, deuterostomy, deubiquitinase, coenocytes, perikymata, rifins;

termed

merotely, bradyopsia, transportins;

2005

called

crenaters, hydrosheds, infrabiological, thencas, marshballs, negadex, lophenteropneusts, boreoeutherian, cardiospheres, supersusceptibility, chronobiotics, argosomes, metalloligands, interchromosome, fruitcases, biseparable, cytonemes, lipopolymer, silicatein, amphipaths, pharyngobasilar, lysobisphosphatidic, rhopalium, pteroid, immunoediting, orexinergic, exoculata, trichoblast, geoneutrinos, transdetermination;

double quotes

anticontributive, methalogical, fluxclimatology, airsurfers, plesiometacarpaler, immunobots, negadex, oxifiers, verticornis, urmetazoan, antagomirs, astrocytogenesis, quantitativity, nonconvecting, hemangiogenesis, proteorhodopsins, biocleaning, picobot, nontronites, cytonemes, femtotechnology, wiregrid, palagonitization, nepotistically, haemangioblast, transactivity, galaninergic, enteropneusts, gigaxonin, hydrinos, microcephalics, finized, palaeoanthropologists;

single quotes

aubotsy, parasuperconducting, mutalecimes, nongouge, adipocrine, nutristad, hydrometropole, nanobaubles, hydrosheds, oscillophor, cotransin, infrabiological, pipmodulins, interologues, composome, neophrenological, pardoides, kilogirl, sequelog, premammilary, metabolator, negadex, lophenteropneusts, edentus, hyopsodontids, hyopsodontid, lithoheterotrophic, acetoxonium, sequelogue, superpigment, lophenteropneust, chemicurrent, pseudodimer, energeticists, nanogratings, antagomirs, supersusceptibility, anthropodenial, omovertebral, megamullion, clusteredness, brainprints, calcichordates, pseudoproxy, ladderane, superphyla, chromathography, magnetoelectrics, finitists, metanodes, prepatterns, cryovolcano, halleyan,

blattellaquinone, repressilator, inchworming, downblended, dewetted, biospherians, destratified, postselected, aggrecanases, interologs, superspreading, condylarth, subfilaments, pterobranchs, lymphohaematopoietic, adakite, condylarths, superspreaders, pangenes, multiferroics, sequenceable, dehydrative, nanoreactor, tsunameters, triallelic, superstrains, stressmeter, segnosaur, drugable, superrotation, pseudospins;

defined as
shuttings;

known as
bonannione, stylocone, chelifores, strigolactones, prosensory, urmetazoan, transresistivity, arogyapacha, retrotranspose, unsynapsed, rhizophore, diacylglycerides, gigaxonin, anthracotheres, androdioecy;

termed
wingbia, metabolator, acetoxonium, chemicurrent, conpats, antagomirs, minisleep, magnetoelectrics, enteropneusts, idiomorphs, deimination, multiferroics, tracheoles.

5 Analysis of results

Table 4 below shows statistical data related to the respective discriminants and the *random* file as a reference (summary for all years).

Table 4. Statistical data of the discriminants

Discriminant	Number of neologisms	Number of neologisms/Number of occurrences of a discriminant	Number of neologisms/Total number of analyzed words
called	163	0.020	0.0015
defined as	16	0.012	0.0006
known as	84	0.020	0.0014
termed	44	0.043	0.0032
double quotes	191	0.004	0.0003
single quotes	581	0.011	0.0061
random	3	-	0.00055
Total	1082	-	-

6 Expandability

The proposed method can easily be applied to other languages with requisite modifications, both with respect to graphical and lexical discriminants.

As concerns lexical discriminants, English, as a mostly analytic language, is relatively easy when it comes to the retrieval of phrases as they are fixed (*termed*, *known as*, *called*, *defined as*) and so can be easily found in a text as they are, without the need to make allowances for inflection.

Fusional languages, in turn, employ inflectional forms, which makes retrieval a little more challenging process, albeit a feasible one if regular expressions are used. Take the Polish language as an example:

the following phrases can be used as discriminants:

1. *tzw*. (so-called).
2. *określan*[a-z]+ *jako* (defined as)[23].
3. *definiowan*[a-z]+ *jako* or *definiuje się jako* (defined as).
4. *zwan*[a-z]+ or *nazywan*[a-z]+ (called).

(The method has been described in detail with respect to most of these discriminants in Wierzchoń 2002.)

Likewise, certain adjustments have to be made with respect to graphical discriminants. Polish uses low left quotes, which have to be converted to ASCII signs ("). Furthermore, single quotes are not used (at least not correctly); therefore, the valuable distinction observed in English is lost and both citations (in English: double quotes) and tentative, ironical, neologism or neosemanticism uses (in English: single quotes) are marked in the same way; therefore, the efficiency of graphical discriminants in the excerption of neologisms (number of neologisms per number of words total) may be lower than in the English language.

7 Further development

The method, albeit highly automated in principle, has been implemented step by step, even though supported by the use of more or less sophisticated programs. Still, it could easily be developed into a complete suite to provide a highly automated tool for the excerption of neologisms. The user would define the input text, specify the phrases (discriminants of neologisms) or punctuation marks and set the threshold. The output would be a list of words ranked according to the number of hits (occurrences) assigned to each. The "only" manual step would be the final analysis of the word list in search of typos, foreign words, ephemeral forms etc. Depending on the availability of morphological analyzers it could even operate on more than one language.

8 Discussion

Even though the method has yielded an extensive set of interesting neologisms, it has certain intrinsic shortcomings:

a. not all words in a text are analyzed, but only those within the graphical discriminants or following the lexical discriminants; however, it is thanks to the use of the discriminants that the likelihood of finding an interesting neologism is far higher than searching at random and faster compared to manual methods;
b. due to the nature of the operations some forms are lost, such as all the capitalized words (e.g. those that follow or precede acronyms or at the beginning of sentences); the latter doesn't apply to lexical discriminants as they are preceded by a set phrase in a sentence;
c. not fully automated, not only with respect to the final manual analysis, but also all the successive steps; with respect to a fully manual approach, however, it seems to be progress.

At the same time the use of the search engine made it possible to automatically eliminate plenty of unwanted words[24] from the lists acquired after morphological analysis, such as (examples from the 1997 year's issue):

[23] Syntax used in regular expressions; + denotes 'one or more occurrences of any sign.' Thus, określan[a-z]+ represents a set that includes all cases, genders, and numbers (określanego, określanej, określanych).
[24] That is, the number of their occurrences in the search engine was higher than the threshold (for example, 1000 hits).

a. typographic errors: *thedynamics, contraints, searchfor, comittee, thepower, c oncerns, worldof, recurr, systemwith, necesssarily, itright.*
b. foreign words: *oeconomicus, stephensi, voor, laboratoire, universités, burgdorferi, conventionné, elegans.*
c. archaic forms: *freind, helpe, onely, yeares, finisht, joineth, maketh.*

As for the efficiency of discriminants, the most productive one is obvious: single quotes, both in terms of absolute numbers (581 words) and the ratio of the number of neologisms to the total number of analyzed words (0.0061). At this point the difference in usage between single and double quotes is seen: Double quotes are used when quoting someone; hence, the average number of words per occurrence, 13.4, as compared to 1.8 for single quotes and efficiency lower than in the case of the *random* file. (Intuitively, the number of neologisms in direct quotations is likely to be lower than in the rest of a scientific text.)

Single quotes, in turn, are expressly used to mark new uses, concepts (‘RNA world’, ‘pocket universes’), and words (‘biospherians’), tentative applications or ironic uses (‘very’). Therefore, their application in the excerption of neologisms in the English language is most justified.

Lexical discriminants are also productive (in particular *termed* and *called* and less so *known as* and *defined as*), especially in terms of the number of neologisms per number of occurrences. It will have to be probed in further studies whether any other productive discriminants exist in the English language and what discriminants should be used in other languages.

Bibliography

Bańko, M. 2001. *Z pogranicza leksykografii i językoznawstwa*. Warszawa: Wydawnictwo Wydziału Polonistyki UW.
Bień, J. S., Szafran, K. 2001. Analiza morfologiczna języka polskiego w praktyce. *Biuletyn Polskiego Towarzystwa Językoznawczego*, LVII, pp. 171-184.
Buczyński, A. 2004. *Pozyskiwanie z Internetu tekstów do badań lingwistycznych*. Warszawa: Instytut Informatyki UW.
Buttler, D. 1962. Neologizm i terminy pokrewne. *Poradnik Językowy*, 5-6. pp. 235-244.
Buttler, D. 1993. Neologizmy z formantem -acja w powojennej polszczyźnie. *Przegląd filologiczny,* 38. pp. 7-15.
Chlebda, W. 1991. *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*. Opole: WSP.
Dias, G. et al. 2000. Normalization of Association Measures for Multiword Lexical Unit Extraction. *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA'2000)*. Monastir, Tunisia. pp. 207-216.
Doroszewski, W. 1958-1969. *Słownik języka polskiego*. Warszawa: Wiedza Powszechna.
Golding, A.R., Schabes, Y. 1996. Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. Santa Cruz, CA.
Gries, S. Th., Stefanowitsch. 2004. A. Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9:1. pp. 97-129.
Krzemińska, W., Nowak, P. (eds). 2002. *Przestrzenie informacji*. Poznań: Sorus.
Moszczyński, R. 2006. *Formal approaches to multiword lexemes*. Warszawa: Instytut Anglistyki UW.
Puppel, S. (ed.). 2005. *Scripta Neophilologica Posnaniensa*. Tom VII. Poznań: Wydział Neofilologii UAM.
Siepmann 2005. Collocation, colligation and encoding dictionaries. Part I: Lexicological Aspects. *International Journal of Lexicography*, 18(4). pp. 409-443.
Smółkowa, T. 2001. *Neologizmy we współczesnej leksyce polskiej*. Kraków: IJP PAN.
Stoberski, Z. 1976. O centralną rejestrację neologizmów naukowych. *Poradnik Językowy,* 4. pp. 186-189.
Wawrzyńczyk, J. 1994. *Tak zwane nowe słownictwo polskie w świetle dokumentacji „Polskiego Informatorium Wyrazowego"*. Katowice: Śląsk.
Wawrzyńczyk, J. 1999. *Nowe słownictwo polskie. Fikcje i fakty*. Warszawa: UW.
Wawrzyńczyk, J. 2000. *Słownik bibliograficzny języka polskiego: wersja przedelektroniczna. T. 1, A-Ć*. Warszawa: Uniwersytet Warszawski. Instytut Informacji Naukowej i Studiów Bibliologicznych.
Wierzchoń, P. 2002. Automatyzacja ekscerpcji definiowanych połączeń wyrazowych. Filtry wyrażeń regularnych. In Krzemińska, W., Nowak, P. (eds.). 2005. *Przestrzenie informacji* (pp. 119-184). Poznań: Sorus.
Wierzchoń, P. 2003. *Z cudzysłowów do poczekalni leksykograficznej.* Warszawa: KLiKR UŁ.
Wierzchoń, P. 2005a. *Z cudzysłowów do poczekalni leksykograficznej II.* Warszawa: KLiKR UŁ.
Wierzchoń, P. 2005b. Automatyczne metody ekscerpcji neologizmów, czyli językoznawstwo faktograficzne. In Puppel, S. (ed.). 2005. *Scripta Neophilologica Posnaniensa.* Tom VII (pp. 221-240). Poznań: Wydział Neofilologii UAM.

I would like to thank my tutor, Prof. Piotr Wierzchoń, PhD, for his invaluable help and comments.