IN SEARCH OF LEXICAL DISCRIMINATORS OF DEFINITION STYLE: COMPARING DICTIONARIES THROUGH N-GRAMS[1]

Mariusz Piotr Kamiński

College of Management 'Edukacja' Wrocław, Poland and University of Applied Sciences Nysa, Poland (mariusz20ski@gmail.com)

Abstract

This study investigates definition style in the range of English dictionaries through the analysis of distribution and frequency of most frequent words (unigrams) and word clusters (trigrams). The main goal of the study is to demonstrate how dictionaries can be compared quantitatively. The texts studied were the following six dictionaries: *Johnson* (1785), *Webster* (1865), *OED* (1888-1928), *Chambers* (1952), *LDOCE* (2005) and *COD* (2011). A selection of definitions were subjected to hierarchical cluster analysis and correspondence analysis. The former analysis confirmed the assumption that recent dictionaries, *LDOCE* and *COD*, are most distant to the other dictionaries with regard to the distribution of unigrams and trigrams. The latter analysis pointed to a few words and word clusters (lexical discriminators) which seemed to be distinguishing features of the dictionaries. It is hoped that the methods demonstrated in this paper will expand an inventory of instruments of lexicographical comparison, and will provide more insights into lexical composition of definitions.

1. Introduction

Like other texts, dictionary definitions vary in style according to the editorial policy, the profile of the target users, the lexicographer's individual choice of words, commercial pressures, and other features that may or may not be specified in the style guide. As a result, two different dictionaries are likely to adopt different lexical, syntactic, and pragmatic choices when defining one and the same entry. In order to capture the style of a text quantitatively, and at the same time to determine how much it is similar to or different from another, we can apply one of the methods of computational stylometry.

Computational stylometry is better known as a set of methods used in identification of disputed authorship. Texts written by the same author are assumed to share certain stylistic features, such as word and sentence length, vocabulary richness, the choice and frequency of words, overlap and distribution of n-grams, and syntactic and collocational preference (Oakes 2009: 1071). Because in natural texts these features are not amenable to conscious control of the author, they provide a useful measure of the author's style, and consequently help identify the most likely author in case of doubt (ibid, 1076). Although definitions are typically written not by one lexicographer but a team of lexicographers and experts following the publisher's guidelines, there is no reason why definitions should not be the object of stylometric research.

[1] This is a post-print version of an article accepted for publication in International Journal of Lexicography following peer review. The version of record of Kamiński, M.P. In search of lexical discriminators of definition style: Comparing dictionaries through n-grams, *International Journal of Lexicography* 29(4), 403-423 is available online at: https://doi.org/10.1093/ijl/ecv038

A definition style has been discussed at greater length in the context of foreign language learners (for example Cowie 2002, Dziemianko & Lew 2013, Lew & Dziemianko 2006, Nesi 2000, Rundell 2008, Stein 1989, Wingate 2002), but the lexical composition of definitions in a wider comparative perspective seems to be an under-researched area. A common method of comparing definitions consists in an analysis of similarities and differences in a restricted selection of corresponding definitions from different dictionaries or editions of the same dictionary (Ilson 1986a, 1986b, Kamińska 2014, Kamiński 2013). Although this method yields valuable insights into underlying lexicographic policies, it does not provide sufficient evidence on how countable features are distributed in the whole dictionary. Such information would usefully supplement what one can glean from a manual comparison of material (see also Coleman & Ogilvie 2009). Therefore, with regard to quantitative data, dictionary comparison can gain more power if it is enriched by a statistical analysis based on a random representative sample, which is capable of confirming or refuting the researcher's initial assumptions.

The current study focuses on the choice and frequency of n-grams, that is words or clusters of adjacent words.[1] N-grams are extracted from a text by capturing the first n-number of words (tokens) in the text, and then moving the n-word window from the beginning to the end of the text, one word at a time (Clough & Gaizauskas 2009: 1257). For example, if we have a sentence *A cat is a small animal*, we can extract the following three-word-strings: *A cat is*, *cat is a*, *is a small*, *a small animal*. If we are interested in extracting one-word strings, we will obtain six of them, as there are six words in the above sentence. N-grams that occur in a given text with a relatively high frequency as compared to other texts will be considered as lexical discriminators. For illustrative purposes, we will take as the units of our analysis unigrams, that is, graphic words (e.g. *cat*, *over-the-top*, *one's*), and trigrams, that is, three-word clusters (e.g. *cat is a*).[2]

The goals of the analysis are, first, to demonstrate how dictionaries can be compared quantitatively; second, to identify dictionaries that are similar with regard to distribution of unigrams and trigrams; third, to identify lexical discriminators, that is, unigrams and trigrams that are most characteristic of each dictionary. The hypothesis is that dictionaries with similar lexical distributions are very likely to belong to the same genre or time period. By identifying lexical discriminators, we hope to highlight lexicographers' preferences for use of certain words and clusters in definitions.

2. Method

2.1. *Materials*. The study was conducted on six well-known dictionaries: Johnson's *Dictionary of the English Language*[3] (1785), Webster's *American Dictionary of the English Language* (1865), *A New English Dictionary on Historical Principles* (1888-1928), *Chambers's Twentieth Century Dictionary* (1952), the *Longman Dictionary of Contemporary English* (2005), and the *Concise Oxford English Dictionary* (2011). In this paper these dictionaries will be referred to as *Johnson*, *Webster*, *OED*[4], *Chambers*, *LDOCE*, and *COD*[5], respectively. All these books are remarkable achievements in English-language lexicography, with *Johnson*, *Webster*, and *OED* being undisputable landmarks. Most dictionaries studied belong to the British tradition of dictionary making, with the exception of *Webster*, which is properly speaking part of American lexicography. While *OED* is scholarly in nature, *Chambers* and *COD* are popular reference books intended for a general audience. *LDOCE* belongs to EFL lexicography, and is addressed to learners of English as a foreign language. Each dictionary differs from another in one or more respects, but what is of interest to us are differences in lexical composition of their definitions.

2.2. *Preparation of data*. The preparation of the source material from each dictionary consisted of several stages. They were as follows:

1) selection of dictionary pages,
2) conversion of page images into text files,
3) extraction of definitions from the files;
4) proofreading and correction of errors in definitions
5) combining data from definition files
6) generating a sample of definitions consisting of ten text fragments of equal length (16000 word tokens)
7) converting the sample of tokens (unigrams) into trigrams.

In the first stage, a simple random sample of pages was selected from each dictionary. Each dictionary sample contained a definition text of comparable length, which spread over 30 to 60 pages, depending on text density in particular dictionaries. The proportions of the sample size to dictionary size, as measured in page numbers, were as follows: 0.3% for *OED*, 1.5% for *LDOCE*, 1.8% for *COD*, 1.9% for *Webster*, 2.3% for *Chambers*, and 2.7% for *Johnson*. [6] No prior decisions were made with regard to the choice of letters from which the pages were to be selected. In order to ensure that every page in a dictionary has an equal chance of being selected, we used an online software, Random.org, to generate random page numbers. It was decided to draw pages rather than entries, as only the former can be randomised easily (cf. Bukowska 2013: 29). It should be pointed out that a comparison of corresponding entries or definitions was not the aim of this study.

In the second stage, the pages from the dictionaries that were available only in hard copies (i.e. *Chambers*, *COD*, *LDOCE*) were scanned first, and then recognised with the aid of OCR software.[7] The other dictionaries (*Johnson*, *Webster*, *OED*) were downloaded from the Internet (Canadian Libraries) as pdf files, and since they did not require scanning, they were processed straightaway.

In the next stage, as soon as the selected pages were read into the OCR software, the definitions were tagged with a view to their retrieving. Tagging was conducted manually while the page was still loaded into the software. Tags were applied in such a way as to mark the whole definition sections rather than individual definitions so as to avoid the problem of where a definition boundary falls. Other dictionary information, such as pronunciation, etymology, illustrative examples, classificatory labels, etc, was discarded. Tagging consisted in inserting a unique symbol, such as a left angle bracket, to mark the beginning of a definition section, and a right one, to mark the end of it, as in the following entries:[8]

> **creek**, *krēk*, <a small inlet or bay, or the tidal estuary of a river : any turn or winding : in America and Australia, a small river or brook.> –adj. **creek′y**, <full of creeks : winding.>
> [Prob. Scand., O.N. *kriki*, a nook; cf. Du. *kreek*, a bay; Fr. *crique*.]

As soon as definitions were tagged, they were generated in the form of text files. For the purpose of preparation of data at this as well as subsequent stages of analysis, I wrote scripts in R[9] (R Development Core Team 2013).

The next step consisted in careful proofreading of definitions and correcting of errors made by the OCR software technology. The aim was to make sure that original spelling was preserved, regardless of orthographic variation, for example, in compounds. Proofreading the older dictionary texts (such as those of *Johnson*, *Webster,* and *OED*) was more demanding than of the recent dictionaries. For example, Johnson follows the older spelling convention according to which the letter “s” is coded by “ʃ” or “ſ”, which required more manual intervention on the part of the proofreader. In addition, the quality of old dictionary copies is generally poorer than of the modern books, and the risk of typos and OCR errors is higher. Nevertheless, proofreading was facilitated by the interface of the software which showed simultaneously in separate windows both the converted text and the original page image, allowing for easy comparison of texts and fast detection of mistranscriptions.

The fifth stage of data preparation aimed at obtaining equal samples of definition texts to be analysed. The idea was to obtain ten samples for each dictionary, as our method of analysis, in particular hierarchical cluster analysis, required that the data be divided into several groups in order to show that they are more similar to one another within a dictionary than between the dictionaries. With this aim in view, the data from definition files were combined into a vector, that is, a sequence, of running tokens. For each dictionary, a vector of 16000 tokens was produced, and then split into ten text fragments of equal length (1600 tokens). As a result, we obtained ten samples of unigrams for each dictionary.

Finally, in order to conduct the research on trigrams, the above samples were converted into trigrams. The samples of both unigrams and trigrams were subjected to further manipulation, depending on the type of the research, that is, hierarchical cluster analysis and correspondence analysis.

2.3. *Data analysis*

2.3.1. *Hierarchical cluster analysis of unigrams and trigrams*. The question of interest to us was whether certain dictionary samples are more similar to each other than to others with regard to the distribution of most frequent unigrams and trigrams. In order to answer this question, we conducted two separate analyses, one for unigrams and the other for trigrams. In the former, we prepared a table of top frequency unigrams in the whole dictionary sample, that is, word types occupying the first 1-1500 ranks. Because we had ten samples of each dictionary, we considered ten counts of word frequency. As a result, we obtained a matrix of 1500 rows and 60 columns (see Table 1), with each row representing a word type, and each cell showing how often this type occurred in a dictionary sample. Words are ordered according to their rank (position) in the frequency table obtained from the whole dictionary sample.

The data from Table 1 were subjected to hierarchical cluster analysis. We generated a cluster dendrogram (see Fig. 1) showing which dictionary sample enters into correlation with other samples with regard to similarity of word distribution.

The methodology of the analysis of trigrams was identical with that of unigrams. It consisted in the preparation of a table of top frequency trigrams, counting their occurrences in each sample, and then submitting the data to a cluster analysis. The results are shown in Fig. 3.

2.3.2. *Correspondence analysis of unigrams and trigrams*. In order to explore in more detail similarities and differences between the dictionaries, we conducted correspondence analysis followed by tests for significance. We addressed the question of which unigrams and trigrams correlate with which dictionaries by virtue of their relative frequency.

The analysis was first run on unigrams, that is words. For each word in a dictionary a mean frequency was calculated, as shown in Table 2. This table is the beginning of a large matrix, with rows representing dictionaries, and columns – words. The latter were ordered according to word rank in a frequency table, from 1 to 14146[10], but only the most frequent words (ranked 1 to 50) were subjected to analysis. Correspondence analysis gave us a pictorial representation of dictionaries and words on the same set of axes (see Fig. 2). In this representation, distances between different data points were interpreted as differences or associations between them. In its essence, correspondence analysis consisted in calculating differences between the rows and columns of frequencies, converting them into distances, and then plotting them graphically. The interpretation of data visualisation in this analysis was supported by a statistical test for significance, which was applied in order to assess whether a difference between word frequencies was statistically significant.

The above study was repeated on trigrams. The data submitted to the analysis were collected in a similar table as the one mentioned above, but this time columns represented trigrams (see Table 4). Although the original table contained 85819[11] trigram types placed in columns, the analysis was conducted on a subset of most frequent trigrams, that is, the first 50 columns. The output of the analysis is shown in Fig. 4 and 5.

3. Results

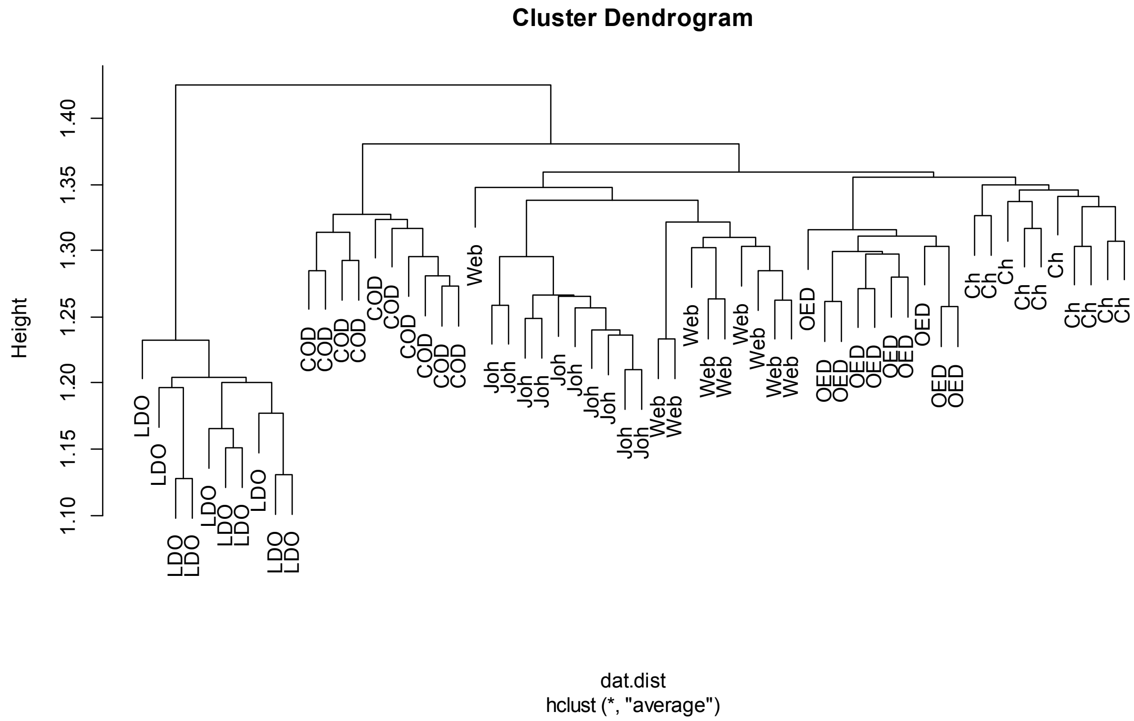
3.1. *Unigrams: Hierarchical cluster analysis*. Table 1 is the beginning of the matrix of word frequencies in each dictionary sample. The first ten columns display frequency counts for words in *Chambers*, followed by *COD* and other dictionaries. As can be seen, words occupying top frequency ranks are function words, which is typical of natural texts. We ran cluster analysis on the first 1500 rows and obtained the results which are shown in Fig. 1.

Table 1. The beginning of the table of frequencies of unigrams in dictionary samples.

rank	*type*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*COD*	*COD*	*COD*	*COD*	*COD*	*COD*
1	**a**	106	158	131	111	134	132	145	171	89	127	142	137	133	131	140	118

2	**of**	103	94	76	97	103	106	93	74	78	83	74	87	106	102	74	83
3	**the**	85	64	77	86	92	84	65	73	60	101	58	61	76	83	58	67
4	**to**	79	84	67	59	58	56	111	45	140	89	35	42	30	35	32	36
5	**or**	60	65	46	43	53	56	57	50	68	60	83	85	82	58	104	76

Figure 1. A dendrogram generated by cluster analysis of unigrams.[12]



As might be expected, most dictionary samples cluster into groups according to their source. At the bottom of the diagram, we can see that texts from the same dictionary are linked to form pairs, and then larger clusters, showing that they enter into stronger correlation with each other than with other dictionary texts. At the highest level of division, *LDOCE* makes a separate group, which suggests that this dictionary has a distinct pattern of word frequencies. The next split isolates another dictionary, that is, *COD* from the rest. These remaining dictionaries fall into two branches. The first branch shows that *Johnson* is close to *Webster*, and the other one that *OED* is close to *Chambers*, with regard to distribution of word frequencies. The similarities among dictionaries are based on a strong correlation between patterns of word distribution.

3.2. *Unigrams: Correspondence analysis*. Table 2 shows part of a matrix of mean frequencies of words across dictionaries. These data were submitted to correspondence analysis in R, the results of which are presented in Fig. 2. As can be seen in the table, words occur with varying frequencies, but what counts in this type of analysis is their relative frequency, calculated as their proportion in rows and columns. These values were used to plot Fig. 2.

Table 2. The beginning of the table of mean frequencies of unigrams in each dictionary.

	a	**of**	**the**	**to**	**or**	**in**	**and**	**an**	**with**	**that**	**by**	**as**	**is**
Ch	130.4	90.7	78.7	78.8	55.8	37	16.2	19	17.1	7	13.6	12.1	3.6
COD	133.8	85.7	65.9	35.8	79.9	35.9	26	20	18.3	7.7	15.7	11.8	8.1
Joh	68.9	78.5	87.8	121.3	36.2	35.1	17.4	12.3	18.6	16.6	19.5	12	22.8
LDO	97.2	52	50.3	67.4	62.2	36.4	25.4	13.2	9.9	40.1	7.3	7.3	29.8
OED	82.9	107.3	88	74.3	80.9	38.7	13.3	16.1	15	8.7	14.2	14.1	6.7
Web	82.6	103.1	99.8	86.8	71.7	36	27	17	14.5	7.2	14.8	26.8	10.5

Figure 2. Correspondence analysis of unigrams ranked 1 to 50.[13]

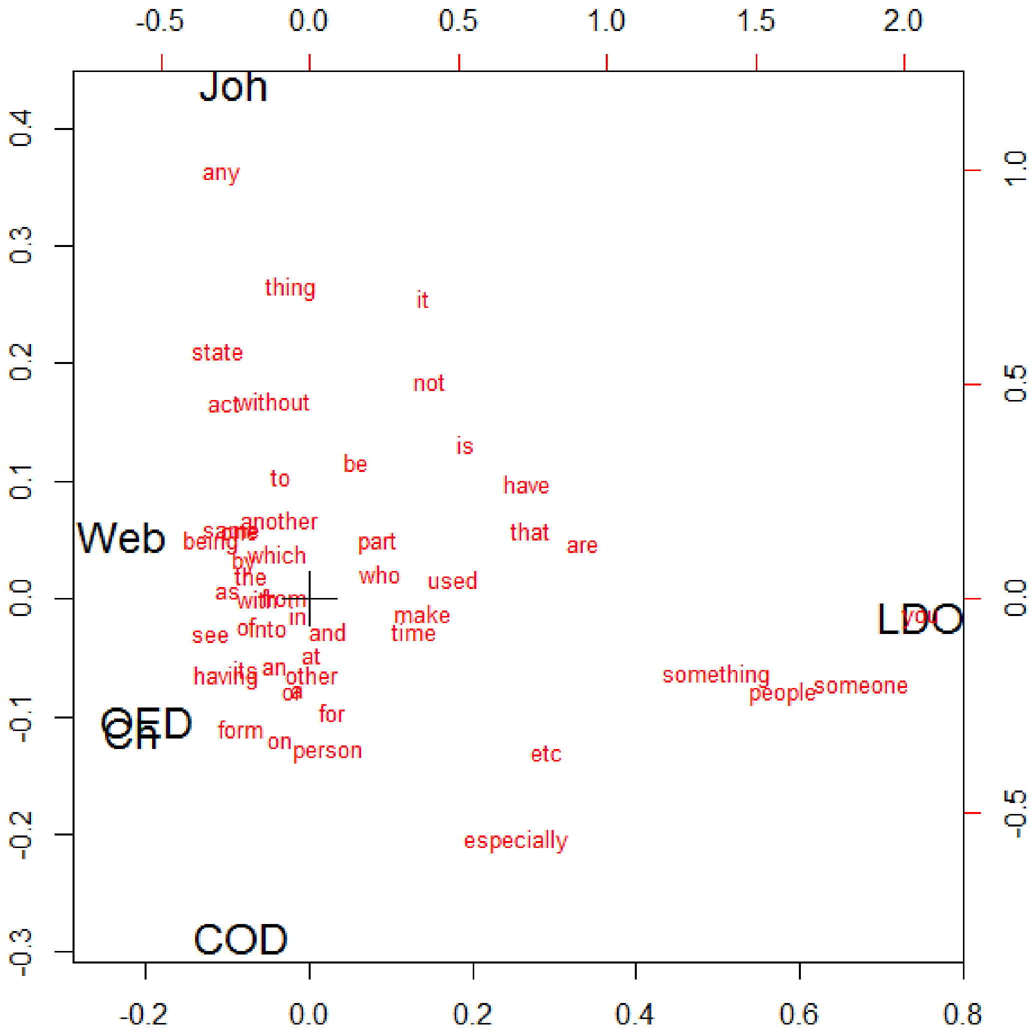


Fig. 2 visualises a subset of most frequent words (ranked 1 to 50) in relation to themselves as well as to the dictionaries. This plot should be interpreted in such a way that data points appearing in proximity to each other show strong association, and the degree of the association is greater for points located away from the centre of the diagram (see the cross), than for those in the proximity to the centre. For example, because a word *someone* has a much higher occurrence in *LDOCE* than in other dictionaries, it appears close to this dictionary, and a long distance away from the centre, showing the association to be strong.

In Fig. 2 the main distinction is visible along a vertical axis, which clearly distinguishes *LDOCE* from the other dictionaries. *LDOCE* is the only dictionary located in the right side of the plot, far from the others. This fact testifies to a huge distance that separates this dictionary from the others with regard to the frequency of words ranked 1-50. It also corresponds well to the distinct status of *LDOCE* as shown in cluster dendrogram in Fig. 1. In

Fig. 2, words correlating with *LDOCE* include *you*, *someone*, *something*, *people*, *especially*, *etc*, *used*, *are*, *that*, *make*. A test for significance indicated that these words occur more frequently in *LDOCE* that in the other dictionaries.[14]

All the remaining dictionaries are distinguished along horizontal axis, which divides the plot into upper and lower quadrants. The former one contains *Johnson* and *Webster*, and the latter *OED*, *Chambers* and *COD*. In the upper quadrant, *Johnson* co-occurs with characteristic words, which are somewhat spread away from the centre, indicating that the association is strong. For example, based on the plot and a significance test[15], we can say that *any*, *thing*, and *without* are used more frequently in Johnson than in the other dictionaries[16], while *state* and *act* are typical of both *Johnson* and *Webster*, the latter dictionary being located in the same quadrant of the plot as the former is. Other words correlating with *Webster* and occurring in this dictionary more frequently than in the other books include *as* and *being*.[17] Respective words in *OED* include *form* and *having*[18]; while in *COD*: *person* and *especially*[19], of which the latter also highly correlates with *LDOCE*. Although *Chambers* is located in the same place as *OED*, it seems to be less marked by lexical preference, as words occurring in this dictionary significantly more frequently than in the other dictionaries are more difficult to find, one such a word being *a*.[20] The rather indistinct composition of Chambers definitions will also be confirmed in the analysis of trigrams (see section 3.4).

3.3. *Trigrams: Hierarchical cluster analysis*. The matrix containing mean frequencies of trigrams is presented in part in Table 3. In order to see which dictionary samples show similar distribution of frequencies, the data from this table were subjected to hierarchical cluster analysis. The results are presented in Fig. 3.

Table 3. Trigram frequencies in each dictionary sample (the beginning of the table).

rank	*type*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*Ch*	*COD*	*COD*
1	**the act of**	1	1	0	1	1	0	1	0	1	0	0	0
2	**part of a**	0	2	3	0	1	0	5	0	0	1	1	1
3	**the state of**	0	0	0	0	0	0	0	0	1	0	1	0
4	**state of being**	0	1	0	1	0	2	1	0	1	1	1	0
5	**one of the**	0	0	0	1	1	4	1	0	0	1	0	0

Figure 3. A dendrogram generated by cluster analysis of trigrams.

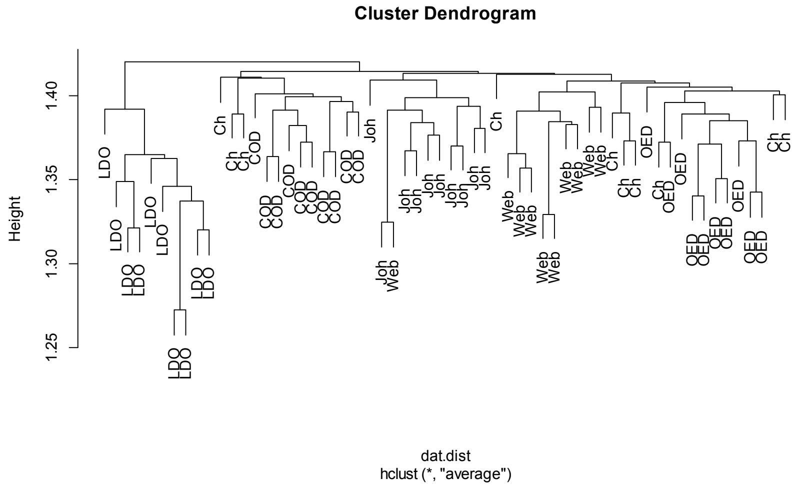


Fig. 3 shows that texts from the same dictionary merge into more or less homogenous clusters, each of them representing a distinct pattern of word frequencies. At the bottom of the figure, we can see that texts from the same dictionary are merged into pairs, and then into larger clusters. As in the analysis of unigrams, *LDOCE* texts form a cluster distinct from other dictionaries. Both *COD* and *OED* clusters are less uniform than in the previous research (on unigrams), as they contain *Chambers* texts. On the whole, however, the results of this analysis are similar to that of unigrams, but the division into clusters is less neat and clear-cut than in the former study.

3.4. *Trigrams: Correspondence analysis*. The data for correspondence analysis are shown in part in Table 4, and the results of the analysis on the most frequent trigrams are presented in Fig. 4 and Fig. 5. The latter figure shows the right side of the plot in close-up.

Table 4. The beginning of the table of mean frequencies of trigrams in dictionaries.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	the act of	**part of a**	**the state of**	**state of being**	**one of the**	**in order to**	**quality of being**	**a kind of**	**the action of**	**a piece of**	**the quality of**	**obs form of**	**pertaining to the**	**of the same**	**in which the**	**or pertaining to**
Ch	0.6	1.2	0.1	0.7	0.8	0.1	0.3	0.8	0.2	0.5	0.1	0	0.5	0.3	0.5	0.4
COD	0	1.1	0.2	0.2	0	1	0.1	0.2	0.8	0.6	0	0	0	0.4	0.5	0
Joh	3.4	0.6	2.4	1.2	0.3	0.1	0.4	1.1	0.2	0.3	0.6	0	0.1	0.8	0.1	0
LDO	1.2	2.1	0.4	0.3	0.7	2.8	0.3	0.3	0	2.1	0.4	0	0	0.2	0.3	0

OED	0.2	0.3	0.1	0.3	1.6	0.3	0.7	0.6	2.9	0.3	0.7	3.7	1.3	0.1	0.8	2.1
Web	4.1	0.5	2.1	2.5	1.3	0.1	2.5	1.2	0.1	0.3	2.1	0	1.3	1.3	0.8	0.5

Figure 4. Correspondence analysis of trigrams ranked 1 to 50.

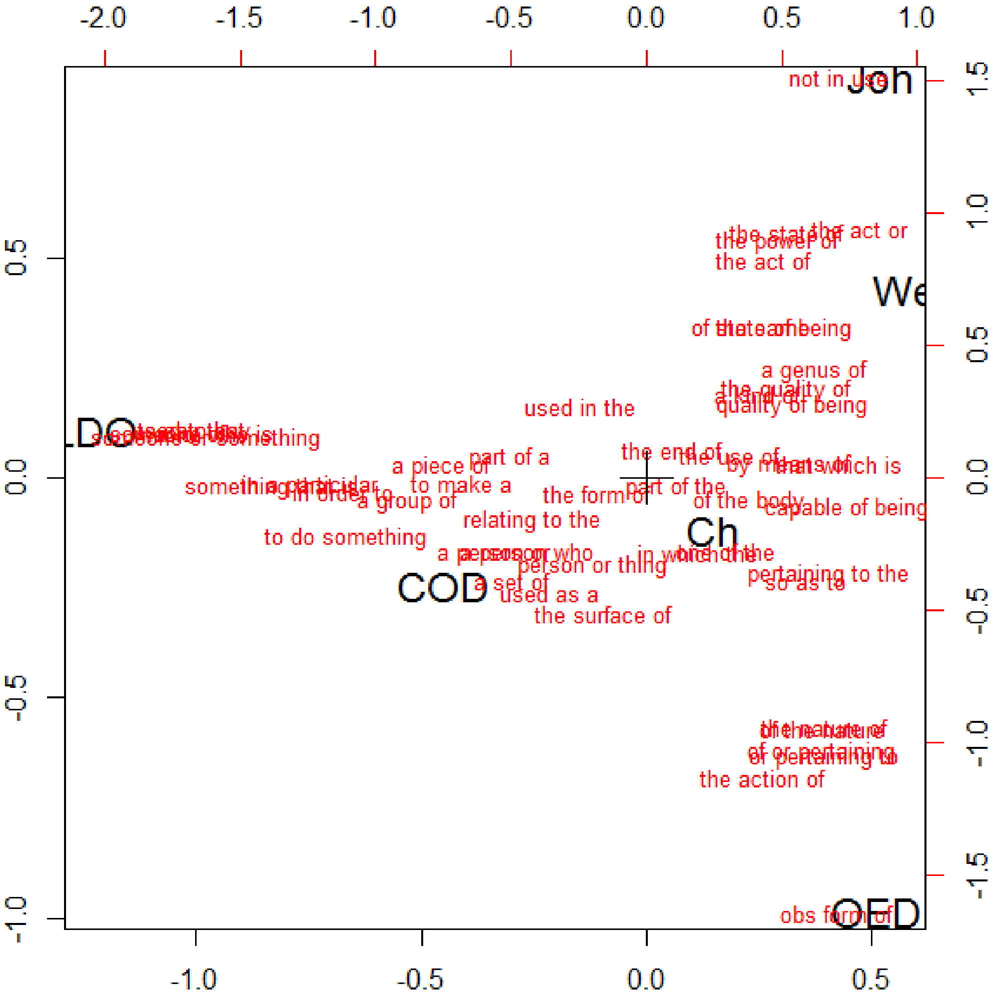
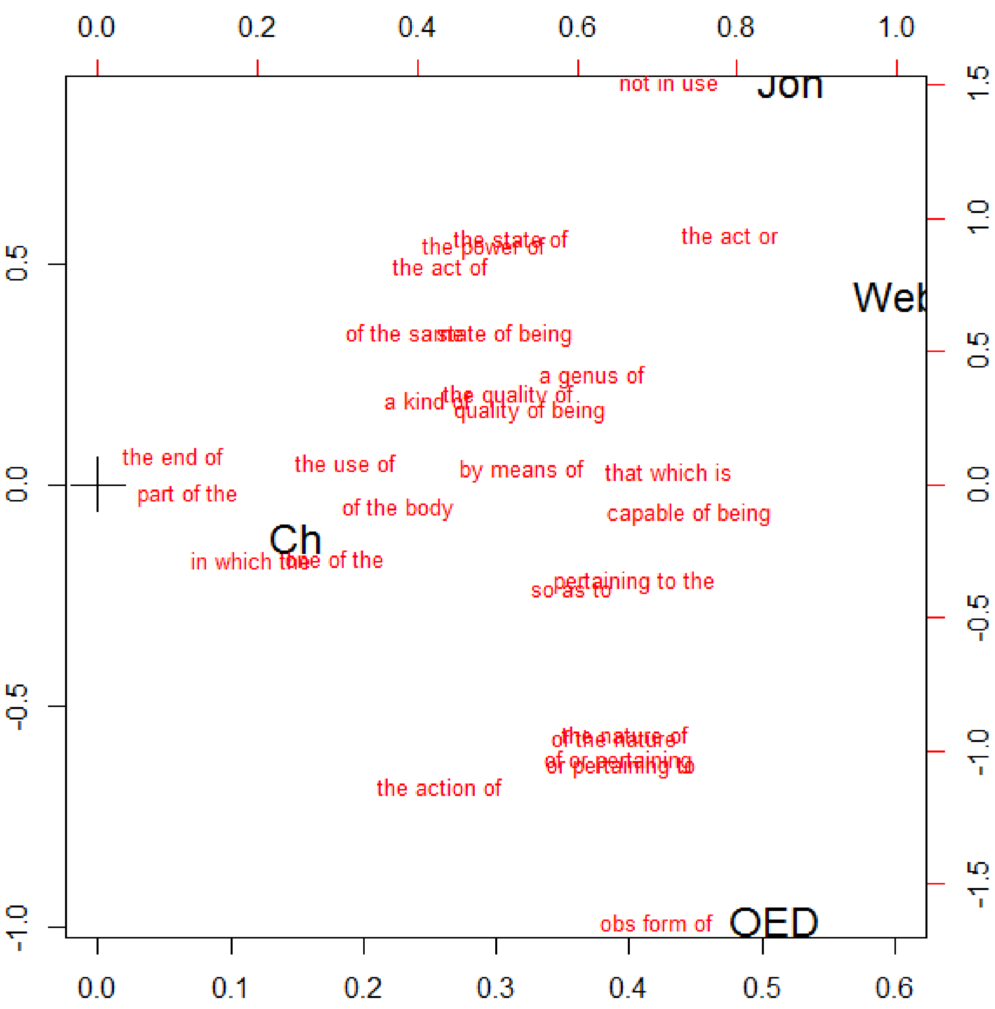


Figure 5. Correspondence analysis of trigrams ranked 1 to 50 in close-up.



As can be seen in Fig. 4 and 5, trigrams spread towards dictionaries, indicating that there are various degrees of correlation between trigrams and dictionaries. The further away from the centre of the plot we move along the axes, a greater degree of correlation we find. In Fig. 4. on the left side of the plot are two most recent dictionaries: *LDOCE* and *COD*. The position of these dictionaries indicates their relative similarity in comparison to the other dictionaries. *LDOCE*, which is placed further away from *COD*, has the following highly distinctive trigrams: *someone or something*, *someone who is*, *to say that*, *used to say*, *a lot of*, *something that is*, *in a particular*, *in order to*, *to do something*, *a group of*, *a piece of*, *to make a*, *part of a*. All of them are used in *LDOCE* significantly more often than in the other dictionaries[21]. The degree of association of points clustered around *COD* is weaker, but there are a few trigrams which are highly characteristic of this dictionary and, compared to the other books, significantly more frequent: "*a person who*", "*used as a*", "*a group of*".[22]

Looking at the right upper quadrant of the plot in Fig. 5, one can see trigrams that are associated with *Johnson*, *Webster* or both. For *Johnson*, the most characteristic combination is "*not in use*"[23], which is in fact a usage note, typographically indistinguishable from definitions. Wordings typical of both dictionaries include *the state of***,** *the act or*, *the power of*, *the act of*, *state of being*, *the quality of*, *the act of*, *the act or*, and *the state of*.[24] Among Webster's favourite trigrams are *a genus of* and *quality of being*, and some other which are located in the right lower quadrant: *pertaining to the*, *that which is*, *capable of being*.[25]

When it comes to *Chambers*, it is the least distinctive book with regard to patterns of use of trigrams, as it is close to the centre of the plot. This finding parallels the one reported on unigrams.

Finally, in the right lower corner of the plot, there are trigrams associated with *OED*. Those which are used significantly more often in this dictionary than in the others are as follows: *obs form of*, *the action of*, *or pertaining to*, *pertaining to the*, *of or pertaining*, *the nature of*, *of the nature*, *so as to*, *capable of being*.[26]

4. Discussion

As was to be expected, hierarchical cluster analysis confirmed the assumption that *LDOCE* differs most from the other dictionaries with regard to the distribution of most frequent unigrams and trigrams. One of the key factors contributing to this finding is the fact that it is the only dictionary to use restricted vocabulary in definitions. As a result of this policy, certain words (see below) occur in this dictionary significantly more frequently than in the other reference books. Cluster analysis also pointed to another dictionary, namely *COD*, as being the second most distant to *Chambers*, *OED*, *Webster* and *Johnson*. The possible explanation of this fact can be sought on historical grounds. Both *LDOCE* and *COD* are most recent dictionaries, published at relatively the same time (in 2005 and 2011, respectively), while the remaining books belong to the rather distant past of the English lexicography, published in different time periods. These historical factors certainly contribute to the shape of the cluster dendrograms shown in Fig. 1 and 5.

The fact that *COD* and *OED* are products of the same publishers is irrelevant, as no stylistic kinship was found between these dictionaries. A more pertinent factor is that they were published at different times, with the intervening period of a century between them. This time gap may be one of the reasons for lack of similarity. Although the editors of the first edition of *COD* (1911) used *OED* consistently as reference[27], they employed a telegraphic style, which was radically different from that used in the parent dictionary. As the editors claimed, *OED* articles were treated as "quarries to be drawn upon than as structures to be reproduced in little" (*COD* 1911: vi). Since then, *COD* has gone through numerous editions and revisions, with the result that it moved even further away from the original style. In 1999 its definitions were thoroughly revised by Judy Pearsall, based on materials and principles used in the compilation of the *New Oxford Dictionary of English* (*NODE*) published a year earlier under the same editorship. Thus, although *COD* is a remote descendant of *OED*, its modern style owes more to *NODE*.

The results of correspondence analysis are more interesting, as they point to specific words and word combinations as distinctive lexical discriminators of definition style in the dictionaries. *LDOCE* stands out from the rest of the dictionaries because of a high degree of correlation of the discriminators with the dictionary (see Fig. 2 and 4). Compared to the other dictionaries, *LDOCE* definitions are characterised by a personal and direct style, which is indicated by a significantly more frequent use of a pronoun *you* and a verb *are*, as in **dim** "if a light dims, or if you dim it, it becomes less bright", and **danger money** "additional money that you are paid for doing dangerous work". Following the innovative style of *Collins Cobuild*, *you* is used in *LDOCE* as an informal way of showing a selection preference of verbs for human subjects (see Hanks 1987, 2006). Furthermore, *LDOCE* is the only dictionary to use in the sample *a lot of*. This expression is another sign of a more user-friendly style, though its more formal alternative *many* is also used in definitions.

Other lexical discriminators for *LDOCE* include unigrams *people*, *someone* and *something*, and trigrams *someone or something***,** *someone who is***,** *something that is***,** *to do something*. Like a pronoun *you*, the above unigrams are used for explanatory purposes. They indicate a selection preference for verbs and make definitions syntactically complete by filling a gap for the subject, as in **grey** "If someone greys, their hair becomes grey", and for verb complements, as in **let** "to allow someone to do something" (*LDOCE*). Such definitions are often more informative than the ones found in traditional native speakers' dictionaries (cf. Rundell 2008); for example, in *COD* the above complementation is left implicit on the assumption that the definition should be substitutable for the word being defined: **let** "allow". Pronouns *someone* and *something* are used for similar purposes in definitions of other lexical categories, such as adjectives, as in **pensive** "thinking a lot about something, especially because you are worried or sad", and nouns, as in "If someone is just a statistic, they are just another example of someone who ..." (*LDOCE*). An obvious advantage of using the above words is that they make definitions intelligible to the non-native user. They belong to a restricted defining vocabulary, which is part of the tradition of vocabulary control movement (see Cowie 2002; Rundell 1998).

Words *someone*, *something* and *people* can also be considered as metalinguistic hedges. If used to express a superordinate concept for the word being defined, they help the lexicographer categorise the word and avoid being overrestrictive. Their combinations such as *someone or something*, *something that is*, form all-embracing categories, indicating hyponymic, or class-based, relations. Trigrams characteristic of *LDOCE* also include such that indicate meronymic (part-whole) relations, specifically: *a group of*, *a piece of*, *part of a*.

In fact, neither of the above types of discriminators, whether hyponymic or meronymic ones, are exclusive to *LDOCE*. Similar combinations also occur frequently in the other dictionaries, for example, in *COD*: *a group of*, in Johnson: *the act of*, *the power of*, *the state of*; in *Webster*: *a genus of*, *the act of*; *OED*: *the action of*, *the nature of*. Arguably, they are typical of the lexicographic style in general. However, it is the relatively frequent use of these expressions throughout the dictionary that makes *LDOCE* distinct from other dictionaries.

While *LDOCE* uses *someone* to categorise the words being defined, *COD* prefers a somewhat formal alternative *person*. This can be easily seen in corresponding definitions of **African** "someone from Africa" (*LDOCE*), and "a person from Africa, ..." (*COD*); and **keyholder** "someone who is officially responsible for keeping the key ..." (*LDOCE*), and "a person who is entrusted with keeping a key ...” (*COD*). Arguably, a number of *person* words entered *COD*, specifically its earlier 10th edition (1999), as part of new definitions taken verbatim from the aforementioned *NODE*, for example the definition for **African** cited above, which replaced the earlier one: “a native of Africa ...” (*COD9*), and the one for **key grip** “the person in a film crew who is in charge of the camera equipment”, which was just added to *COD10* as a new one.

Among other word combinations characteristic of *LDOCE* are *in order to*, which is arguably a manifestation of a more explanatory definition style; as well as *used to say* and *to say that*, which indicate metalinguistic definitions, as in **let me think** “used to say that you need time to think ...”. Words which are highly distinctive of this dictionary include *especially* and *etc*, though again these are typical of lexicographese in general. It is worth noting that the former unigram is also highly characteristic of *COD*. Being a marker of prototypicality, *especially* is significantly more often used in *COD* and *LDOCE* than in the other dictionaries, which may testify to the conscious application of the theory of prototypes in modern English lexicography.[28] This word is one of the features that *COD* has in common with *LDOCE*, and which makes both dictionaries located on the same side of the plot in Fig. 4., in spite of the fact that the dictionaries are intended for different types of audience.

Other dictionaries under study have their own discriminators. *Webster* and *Johnson* are comparable in respect of the use of superordinates, or category terms for the words being defined. Two such lexical items are *act* and *state*, and their combinations: *the act of*, *the act or*, *the state of*, and *state of being*. Other trigrams typical of both dictionaries are *the power of*, and *the quality of*. In turn, *Webster* shows a marked preference for *quality of being*, *that which is*, and *a genus of*. Of these three, the last one is indicative of definitions of plants and animals which are copiously explained in this dictionary. As is seen, *Webster* and *Johnson* share certain highly frequent unigrams and trigrams which are used for classificatory purposes. This finding is consistent with the fact that Webster drew heavily on Johnson when compiling his early dictionary of 1806 (Landau 2005) as well as 1828 (Hanks 2005), of which subsequent revisions led to Webster’s dictionary of 1864.

Unlike in *Johnson* and *Webster*, it is difficult to find discriminators common for *Chambers* and *COD*. This is caused by an incomparable degree of lexical variation across these dictionaries, which can be seen in their position in the plot (Fig. 4). Firstly, *Chambers* is close to the centre, which makes it rather indistinct with regard to lexical preference. Secondly, the two dictionaries are located on the opposite sides of the plot, with *Chambers* being on the right, and *COD* on the left together with *LDOCE*. Moreover, although *COD* has two lexical discriminators *especially* and *person*, none of them occurs in *Chambers* with a comparable frequency. It may be that different publication dates of these dictionaries, namely 1952 for *Chambers* and 2011 for *COD*, lie behind the different patterns of lexical variation.

Trigrams *pertaining to the* and *capable of being*, which are highly distinctive of *Webster*’s definitions of adjectives, are also found in *OED* in relatively large numbers. These *OED* discriminators may be a trace of *Webster*, which was used regularly by Murray as a source material for compilation of *OED* (Silva 2002: 79). Apart from the above mentioned *pertaining to the*, other trigrams specific to *OED* and containing a word *pertaining* include *or pertaining to* and *of or pertaining*. This dictionary also features a relatively high number of *form* and *obs form of*, which are arguably part of numerous metalinguistic definitions of obsolete forms, for example: **Bastailye** “obs. form of Bastille”, **Dedur** “obs. form of Didder”, **Glace** “obs. form of Glass”.

The foregoing analysis was performed on well-known and relatively well-researched dictionaries. Although some of the findings, especially those pertaining to lexical structure of definitions in *LDOCE*, may not be particularly surprising to those familiar with EFL definition style, the current study provides objective, quantitative evidence for what has previously been described on the basis of manual comparisons or intuition. Thus, this type of study may serve as an objective diagnostic of definition style and the lexical composition of definitions.

In order to obtain a more complete picture of stylistic differences, correspondence analysis can be extended to study other features, such as definition length, vocabulary richness, and use of abbreviations or hedges. It is worth trying to study the above features using other types of multivariate analysis, in particular factor analysis, which has been extensively used in studies of genre variation (for example Biber 1995). Furthermore, it may be interesting to investigate user-generated open dictionaries, the content of which is published online and customised in real time by the users. Such a study may reveal the users' underlying lexical preferences.

Notes

[1] N-grams are also termed “lexical bundles”, “clusters”, and “chunks” (Cheng 2012: 7).

[2] We decided to study unigrams and trigrams in order to illustrate the method. However, in research on genre variation, typically it is four-word clusters (i.e. tetragrams) that are taken as units of analysis, as they seem to offer a more easily recognisable range of structures and functions than three-word clusters (Goźdź-Roszkowski 2011: 110, Grabowski 2015: 131).

[3] The choice of the edition was determined by the availability of dictionaries in Canadian Libraries.

[4] Although this dictionary was published under a different title, the *New English Dictionary*, in this paper we use the title *Oxford English Dictionary (OED),* by which the book is now known.

[5] We use an abbreviation *COD*, rather than *COED* representing the recent title, as for a century the dictionary has been widely known under the former name.

[6] The texts analysed, which constituted a small fragment of the whole dictionaries, were extracted solely for the purpose of computational analysis for non-commercial research and are not published in this paper.

[7] OCR was carried out with the use of ABBYY FineReader 11.

[8] Before tagging, it was necessary to make sure that the above brackets were not already present in the dictionary text; otherwise, we would have received false hits. If this was the case, the symbols had to be deleted first.

[9] R is an open source programming language.

[10] This is the number of word types in the whole sample of definitions.

[11] This is the number of trigram types in the whole corpus.

[12] The cluster analysis was performed using R function *hclust*, with the agglomeration method for clustering being “average”.

[13] This plot was obtained in R using *corresp* function from the “MASS” package (Venables & Ripley, 2002).

[14]According to Wilcoxon rank-sum test, words *you*, *someone*, *something*, *people*, *etc*, *used*, *are*, *that* were used in *LDOCE* statistically more frequently (p<0.05) than in all the other dictionaries. A statistically significant result was also noted for *especially*, compared to *Chambers*, *Johnson*, *OED*, and *Webster*; and for *make*, compared to all the dictionaries but *OED*.

[15]According to the test for significance, *act* and *state* were significantly more frequent in Johnson and *Webster* than in all the other dictionaries (p<0.05). A significant difference was obtained for *act* in *Webster*, as compared to *COD*, *LDOCE*, *OED*; and for *state*, as compared to *Chambers*, *COD*, *LDOCE*, and *OED*.

[16] The relatively high frequency of *any* and *thing* in Johnson may be related to the fact that the lexicographer tended to use the spelling form "any thing" rather than "anything" in definitions (Rosamund Moon, private communication).

[17] *as* was significantly more frequent in *Webster* than in all the other dictionaries; *being* was used more frequently in this dictionary than in all other books but *OED*.

[18] A significant difference was obtained for *form* in *OED*, as compared to all the other dictionaries; and *having*, as compared to *Chambers*, *Johnson*, *LDOCE*, and *Webster*.

[19] *person* was more frequent in *COD* than in *Chambers*, *Johnson*, *Webster*; and *especially*, more frequent than in *Chambers*, *Johnson*, *OED*, *Webster*.

[20] *a* occurred in *Chambers* more frequently than in all the other dictionaries but *COD*.

[21] According to Wilcoxon rank-sum test, *someone or something* was used more frequently (p<0.05) in *LDOCE* than in *Chambers*, *Johnson*, *OED*, and *Webster*. This test also pointed to other trigrams as being statistically more frequent in *LDOCE* than in the other dictionaries; these are as follows: *someone who is*, *to say that*, *to say that*, *a lot of*, *something that is*, *to do something*, *in order to*, *a piece of*, *to make a*, as compared to all the other dictionaries; *in a particular*, as compared to *Chambers*, *OED*, *Webster*; *a group of*, as compared to *Johnson* and *OED*; *part of a*, as compared to *Johnson*, *OED*, and *Webster*.

[22] According to Wilcoxon rank-sum test, a significant difference (p<0.05) was observed in use of *a person who* in *COD*, as compared to all the other dictionaries; *used as a*, as compared to *Chambers*, *Johnson*, *LDOCE*, and *Webster*; and *a group of*, as compared to *Chambers*, *Johnson*, *OED* and *Webster*.

[23] *not is use* was statistically more frequent (p<0.05) in *Johnson* than in all the other dictionaries.

[24] A significant difference was noted between the frequency of occurrence of *the power of*, *the act or*, *the act of*, *the state of* in *Johnson* and the corresponding trigrams in all the dictionaries but *Webster*. *state of being* was more frequent in *Webster* than in all the other dictionaries but *Johnson*. Other frequently occurring trigrams in Webster included *quality of being*, as compared to all the other dictionaries but *OED*; *the quality of*, as compared to *Chambers*, *COD*, and *LDOCE*; *a genus of*, compared to *COD*, *Johnson*, *LDOCE*, and *OED*; *the act of*, as compared to *Chambers*, *COD*, and *OED*; *the act or* and *the state of*, as compared to *Chambers*, *COD*, *LDOCE*, and *OED*.

[25] *pertaining to the* was significantly more often used in *Webster* than in *COD*, *Johnson* and *LDOCE*; other trigrams occurring in *Webster* with a higher frequency included *that which is*, compared to *COD* and *LDOCE*; and *capable of being*, compared to *Chambers*, *COD*, *Johnson*, and *LDOCE*.

[26] The test for significance pointed to the following trigrams as being more frequent in *OED* than in other dictionaries: *obs form of*, *the action of*, *or pertaining to*, *the nature of*, and *of the nature*, as compared to all the other dictionaries; *pertaining to the*, compared to *COD*, *Johnson*, and *LDOCE*; *of or pertaining*, compared to *COD*, *Johnson*, *LDOCE* and *Webster*; *so as to*, compared to *Chambers* and *Johnson*; *capable of being*, compared to *COD*, *Johnson* and *LDOCE*.

[27] The Fowler brothers, the *COD* editors, used *OED* as far as it was completed, that is, up to the letter R (*COD* 1911: vi).

[28] Expressions such as *especially*, *typically*, *usually*, *e.g* have long been used by dictionary makers in a more or less regular way, but it is only recently that they have received more attention by lexicographers, who seem to apply the theory of prototypes more consciously and consistently. According to this theory, the expressions can be considered as introducing prototypical features or instances of the category being defined (Geeraerts 2001, 2003).

References

A. Dictionaries

Fowler, H. W., and F. G. Fowler. 1911. *The Concise Oxford Dictionary of Current English.* (First edition). Oxford: Clarendon Press. (COD1)

Geddie, W. 1952. *Chambers's Twentieth Century Dictionary.* (Second edition). Edinburgh: W. & R. Chambers. (Chambers)

Johnson, S. 1785. *A Dictionary of the English Language*. Vol. 1 and 2. London. (Johnson)

Murray, J. A. H. 1888-1928. *A New English Dictionary on Historical Principles*. Vol. 1–10.

(Edited by James Murray, Henry Bradly, William A. Craigie, and Charles. T. Onions). Oxford: The Clarendon Press. (OED)

Pearsall, J. 1998. *New Oxford Dictionary of English*. Oxford: Oxford University Press. (NODE)

Pearsall, J. 1999. *Concise Oxford English Dictionary.* (Tenth edition). Oxford: Oxford University Press. (COD10)

Sinclair, J. 1987. *Collins Cobuild English Language Dictionary*. (First edition). London, Glasgow: Collins. (Collins Cobuild)

Stevenson, A., and M. Waite. 2011. *Concise Oxford English Dictionary.* (Twelfth edition). Oxford: Oxford University Press. (COD)

Summers, D. 2005. *Longman Dictionary of Contemporary English*. (Fourth edition with Writing Assistant). Harlow: Longman. (LDOCE)

Thompson, D. 1995. *The Concise Oxford Dictionary of Current English*. (Ninth edition). Oxford: Oxford University Press. (COD9)

Webster, N. 1865. *An American Dictionary of the English Language*. (revised by C. A. Goodrich and N. Porter). Springfield: G. & C. Merriam. (Webster)

B. Other literature

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Bukowska, A. 2013. 'Sampling in Historical Lexicographic Research.' In *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3).*, edited by R. McConchie, T. Juvonen, M. Kaunisto, M. Nevala, and J. Tyrkkö, 27–34. Somerville, MA: Cascadilla Proceedings Project.

Canadian Libraries. https://archive.org/details/toronto. (Accessed on 10 January 2014)

Cheng, W. 2012. *Exploring Corpus Linguistics*. London and New York: Routledge.

Clough, P., and R. Gaizauskas. 2009. 'Corpora and Text Re-Use.' In *Corpus Linguistics: An International Handbook*, edited by A. Lüdeling and M. Kytö, 2:1249–71. Berlin: Walter de Gruyter.

Coleman, J., and S. Ogilvie. 2009. 'Forensic Dictionary Analysis: Principles and Practice', International Journal of Lexicography, 22 (1): 1–22.

Cowie, A. P. 2002. *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.

Dziemianko, A., and R. Lew. 2013. 'When-Definitions Revisited' *International Journal of Lexicography*, 26 (2): 154–75.

Geeraerts, D. 2001. 'The Definitional Practice of Dictionaries and the Cognitive Semantic Conception of Polysemy' *Lexicographica*, 17: 16–21.

Geeraerts, D. 2003. 'Meaning and Definition' In *A Practical Guide to Lexicography*, edited by P. van Sterkenburg, 83–93. Amsterdam: John Benjamins.

Goźdź-Roszkowski, S. 2011. *Patterns of Linguistic Variation in American Legal English.*

Frankfurt am Main: Peter Lang.

Grabowski, Ł. 2015. *Phraseology in English Pharmaceutical Discourse*. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Hanks, P. 1987. 'Definitions and Explanations'. In *Looking Up. An Account of the COBUILD Project*, edited by J. M. Sinclair, 123–36. London: Collins ELT.

Hanks, P. 2005. 'Johnson and Modern Lexicography'. *International Journal of Lexicography*, 18 (2): 243–66.

Hanks, P. 2006. 'English Lexicography'. *Encyclopedia of Language and Linguistics*. Oxford: Elsevier.

Ilson, R. 1986a. 'British and American Lexicography'. In *Lexicography An Emerging International Profession*, edited by R. Ilson, 51–71. Manchester: Manchester University Press.

Ilson, R. 1986b. 'Lexicographic Archeology: Comparing Dictionaries of the Same Family'. In *The History of Lexicography*, edited by R. R. K. Hartmann, 127–36. Amsterdam: John Benjamins Publishing Company.

Kamińska, M. 2014. *A History of the Concise Oxford Dictionary*. Frankfurt am Main: Peter Lang.

Kamiński, M. 2013. *A History of the Chambers Dictionary*. Berlin: Walter de Gruyter.

Landau, S. 2005. "Johnson's Influence on Webster and Worcester in Early American Lexicography," International Journal of Lexicography, 18 (2): 217–29.

Lew, R., and A. Dziemianko. 2006. 'A New Type of Folk-Inspired Definition in English Monolingual Learners' Dictionaries and Its Usefulness for Conveying Syntactic Information' *International Journal of Lexicography*, 19 (3): 225–42.

Nesi, H. 2000. *The Use and Abuse of Learners' Dictionaries*. Lexicographica Series Maior 98. Tübingen: Max Niemeyer.

Oakes, M. P. 2009. 'Corpus Linguistics and Stylometry' In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 2:1070–91. Berlin: Walter de Gruyter.

Random.org**.** Accessed on 8 January 2015. https://www.random.org/.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria.

Rundell, M. 1998. 'Recent Trends in English Pedagogical Lexicography' *International Journal of Lexicography*, 11 (4): 315–42.

Rundell, M. 2008. 'More Than One Way to Skin a Cat: Why Full-Sentence Definitions Have Not Been Universally Adopted.' In *Practical Lexicography*, edited by T. Fontenelle, 197–210. Oxford: Oxford University Press.

Silva, P. 2002. 'Time and Meaning: Sense and Definition in the OED.' In *Lexicography and the OED*, edited by Linda Mugglestone, 77–95. Oxford: Oxford University Press.

Stein, G. 1989. 'Recent Developments in EFL Dictionaries.' In *Learners' Dictionaries: State of the Art*, edited by M. L. Tickoo, 10–42. Singapore: SEAMEO.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

Wingate, U. 2002. *The Effectiveness of Different Learner Dictionaries: An Investigation into the Use of Dictionaries for Reading Comprehension by Intermediate Learner of*

German. Lexicographica. Series Maior 112. Tübingen: Max Niemeyer.