English Studies at NBU, 2015
Vol. 1, Issue 2, 5-26
ISSN 2367-5705 (Print)
www.esnbu.org

AN ALTERNATIVE PROPOSAL FOR ELICITING KEY WORDS

Elena Tarasheva

New Bulgarian University, Sofia, Bulgaria

Abstract

The article reports research on the concept of key words as statistically significant items in a text or corpus. It reviews approaches to eliciting key words used in various software products for language analysis and the rationale for adopting them. Based on empirical data, a new method is proposed and tested on an exploratory corpus. The motivation and arguments for proposing the procedure are revealed, using comparisons between different languages. The adequacy of the results yielded by the different methods is tested via a mechanism developed with this research.

***Key words*:** corpora, key words, chi-square, log likelihood, lemmas, lemmatization.

Article history:
Received: 22 November 2015;
Reviewed: 14 December 2015;
Accepted: 21 December 2015;
Published: 31 December 2015

Elena Tarasheva is Associate Professor of Discourse Analysis (Media) at the English Studies Department, New Bulgarian University. She obtained her BA in English Philology from Veliko Turnovo University and specialized in Media and Culture Studies at the University of Strathclyde in Glasgow, UK. Tarasheva has her PhD in Mathematical Linguistics from the Bulgarian Academy of Science. She has published two monographs: *Repetitions of Word Forms: an Approach to Text Structure* CSP 2011 and *The Image of a Country created by International Media: The Case of Bulgaria* CSP 2014 and several articles about Cultural Studies, Corpus Linguistics and Political Linguistics.

Email: etarasheva@nbu.bg

In a rare monograph Phillips (1989:11) observed:

[A] distributional analysis of textual substance invoking no knowledge of the semantic content, the syntactic organisation or the lexical meaning of the text would reveal the existence of global patternings in the lexis of the text. […] What the text is about may be specified by providing a semantic interpretation for the formally identified macrostructure.

Since then many researchers have been fascinated by the idea that the lexical structure of texts should be indicative of something bigger. In linguistic circles it is often hinted that corpus-extracted keywords were something that John Sinclair talked about at length, influenced by Phillips' thesis, but published nothing about. Several methods have been introduced of deriving items of key significance and this research purports to contribute to those.

Definitions

In a review of literature on key words, Stubbs (2010: 25) traces them back to Firth's "sociologically important words, which one might call focal or pivotal words". Then he refers to a range of German research, including Teubert's *politische Vexierwört*, which reflect layers of political meanings on the surface and below it. Finally, Stubbs mentions French *mots clés*, embracing Benveniste's concept of civilization. Stubb's coveted goal – also revealed with the title of one of his books (Stubbs 1996) - are key words as indicators of cultural values in society. In this he continues a tradition established with William's list (1976/1983) of culturally significant items - "a vocabulary of culture and society". "Keywords are the tips of icebergs: pointers to complex lexical objects which represent the shared beliefs and values of a culture." (Stubbs 2010: 23).

Baker's definition (2004:350) forges a connection with discourses: "keywords will direct the researcher to important concepts in a text (in relation to other texts) that may help to highlight the existence of types of (embedded) discourse or ideology." While the term 'discourse' has multiple meanings, Baker (2006:2) uses it to refer to a 'system of statements which constructs an object'.

Sinclair (1996) collates cultural significance with textual role: "Keywords are words which are claimed to have a special status, either because they express important

evaluative social meanings, or because they play a special role in a text or text-type. From a linguistic point of view, they contribute to the long search for units of meaning".

The creator of one of the most popular software products for linguistic analysis Wordsmith, (Scott 2001:48) describes keywords via their frequency: "The idea is quite simple: if a word is found to be much more frequent in one individual text than its frequency in a reference corpus would suggest, it is probably a "key word".

In this definition the ambiguity transpires whether we search higher frequency within a text, or in a corpus. We believe there should be a difference between the two, but so far this issue remains unexplored in Corpus Linguistics.

For the purposes of this research, we choose to focus on statistically established words that have a predominance in a corpus. Whether they project cultural values, or textual properties remains to be checked for each particular case. We believe that the role statistically predominant words play is an effect, rather than a starting point in searching for key words.

Methods of eliciting key words

Scott and Tribble (2006: 57) base their approach to establishing key words on repetitive reference. If a proposition – as suggested by Kintsch and van Dijk (1978) – or a sentence – as suggested by Hoey (1991) – is referred to repetitively, then it should have more importance about the text as a whole. Then, Scott and Tribble select a unit to trace that is immediately obvious and straightforward to establish – the word form, without considering any grammatical or lexical suffixes added to it. In the belief that if a concept is referred to more frequently, then it must lead to the basic conceptual load in the text, they look for lexical repetitions. They then establish statistical procedures comparing the percentage of the entire text that this word presents to the percentage the same word presents in a big general corpus.

Further, some languages have inflections and each verb can occur in a number of inflected forms, as is the case with French, for instance. Languages which have cases contain a range of forms for the nouns and adjectives as well. Yet others agglutinate forms. Thus the frequencies depend heavily on the number of inflected forms. This is reflected in the respective frequencies, as Philip (2010:186) rightly observes:

"…. the calculation of key words is dependent on frequency measures and repetition, yet these matters are not entirely unproblematic. In particular, a language with very few inflected forms has more recurrent forms than a fully inflected one, which in turn has fewer forms than agglutinative or infixing languages. While each word form attracts its own distinctive patterning, the dispersion of closely-related meanings over variant forms of a lemma may affect frequency measures and statistical calculations."

Utka (2004) in his analysis of keywords in George Orwell's 1984, lemmatises noun forms in the text, and calculates keywords based on the frequencies of lemmas, rather than individual word forms. Baker (2004) observes that carrying out such a strategy on his corpus of gay and lesbian narratives "would have enabled a more inclusive form of analysis as it most likely would have resulted in the lemma SESSION being key rather than just the word SESSIONS. However, a lemma-based analysis may not always be a useful strategy as particular word forms can contain specific collocations or senses which would be lost when combining word forms together." Thus, working with un-lemmatised corpora seems to have established itself as the standard.

If lexical recurrence is to be interpreted, then serious statistical procedures need to prove that the numbers are not haphazard. Several have been evolved. This research puts forward a tentative suggestion for another one, while trying to check the outcome of existing ones.

The Chi square list compares the frequency of occurrence found experimentally with those expected on the basis of some theoretical model (Oakes 1998:24). In the case where there is no difference between the reference corpus and the target, the null hypothesis applies. The observed value is denoted with O, and the one in the reference corpus – E. The value of O - E is found and squared to give more weight to the cases where the mismatch between O and E is greatest. Thus, the formula is this:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Chi-square can also serve as a measure of evenness of distribution. Equiprobable distributions are characterised by the same chi-square value.

Alternatively, Dunning's log likelihood measure shows if a word or phrase is overused or underused in a specialised corpus compared with a corpus of Standard English. The formula is this:

$$G^2 = 2\sum x_{\mathrm{y}}(log_4 x_4 - log_{\mathrm{e}} m_{\mathrm{e}})$$

where x_{ij} are the data cell frequencies, m_y are the model cell frequencies, log_e represents the logarithm to the base *e*, and the summation is carried out over all the cells in the table (Oakes, 1998, p 42).

Kilgarriff (1996), having compared the chi-square and log-likelihood (also known as G-square) measures, preferred the G-square. Dunning (1993) points out that most vocabulary items are rare, and thus words in the text are not normally distributed. The advantage of the G-square or log likelihood measure is that it does not assume the normal distribution.

In his on-line software for parsing a range of corpora, Davies (2004) uses the log-likelihood calculation for eliciting keywords. Instead of using a reference corpus for his comparison, however, he employs projections – an expected value based on what has occurred so far in the text.

A Proposal for eliciting key words

The proposal proceeds from observations that concepts which are central to a text are usually named with an extended lemma of the respective lexical item. This is particularly true of languages such as Bulgarian, where the articles are bound morphemes and form new items in the lemma. A study (Anonymous 2011) reveals that research articles contain a chain of words which include the singular and the plural form of a word. They are used for giving examples and present the operative items in the research. All the articles in the corpus contain such repetition chains, irrespective of the genre, topic or subject field. Examples are given in Table 1.

Article 1	***Article 2***	***Article 3***	***Article 4***
transition 70. transitions 70. shift 30. shifts 24. utterances 39. utterance 29.	areas 39. area 25. neuron 13. neurons 39.	agreements 37. agreement 32. state 14. states 31. form 13. forms 12.	systems 34. system 32. grammar 62. grammars 16. structures 20. structure 44.
Article 5	***Article 6***	***Article 7***	***Article 8***
indicator 9. indicators 9.	words 33. word 15.	universities 49. university 31.	symbols 38. symbol 13.

paragraph 10. paragraphs 6. sentence 24. sentences 10.	pairs 8. pair 12. contrast 10. contrasts 6. stimulus stimuli		system 28. systems 20. machines 20. machine 32.

Table 1. *Illustrative chains in research articles.*

At the same time, the central concepts in the articles are presented with repetition chains in which the term is repeated in different forms thus allowing the use of different types of reference - generic, specific, classificatory etc. Below is the concordance of the word *fact*, illustrating that the word occurs with the definite, indefinite and zero article:

be taken as a ***brute*** ***fact*** *wrpl*
are a matter of ***brute*** ***fact***
by ***brute*** ***fact*** *i understand kripke to mean*
to explain beyond the ***brute*** ***fact*** *of agreement of responses that*
collective ***fact*** *as solution after concluding that*
found in the ***collective*** ***fact*** *of the agreements in judgment*
to individuals to a ***collective*** ***fact*** *that is observed as*
same situation to the ***collective*** ***fact*** *which is that members of*
not simply describe the ***individual*** ***fact*** *of jones's supposed conformity*
get us from the ***individual*** ***fact*** *that jones is behaving in*
be found in the ***individual*** ***fact*** *of those states of the*
still is no ***internal*** ***fact*** *of the matter to consider*
apparently is no non-regressible ***internal*** ***fact*** *about the purported rule-follower*

Similar patterns are established for items central to short stories and political speeches thus suggesting the conclusion that concepts central to a text appear in different forms of the word.

Even more visibility is provided through corpora of texts in Bulgarian due to the fact that verbs have an extended list of forms inflected for person, number, tense and aspect and nouns can be plural and singular, with the definite, indefinite or zero article.

In a corpus from the Hansards from the Bulgarian Parliament, the speeches of the Prime Minister contain the following list of cognate words:

Победа (victory)
Победените (the beaten)
Победил (I have won)
Победили (We have won)
Победите (the victories)
Победител (the victor)
Победители (the victors)

They are all from the same root in Bulgarian, some are verbs, others – nouns, in different forms. The availability of such a wide range of cognate words indicates a significant interest in the topic on the part of the speaker.

That is why we believe that the list of words of key significance in a text or corpus can be compiled exploring the words which appear with an extended lemma. The fact that the speaker included in his speech several different forms of a word should signal greater attention paid to a topic. Our examples lead us to believe that immediate candidates for inclusion in such analysis are the forms from the grammatical paradigm of a word – the plural and the singular forms of nouns, the inflected forms of the verbs, the case forms of nouns etc. Other members of the key word list would be cognate words – verbs formed from nominal roots and vice versa, as well as other lexical derivatives.

Method and procedure

To test the adequacy of the proposed method for deriving key words, a corpus is compiled. Four types of key word lists are derived from the corpus:

1. The typical chi-squared list derived automatically via the software Wordsmith tools (Scott 2012);
2. The typical log-likelihood list derived automatically via the software Wordsmith tools;
3. The frequency list for the corpus purged of the grammatical high-frequency words;
4. The list of words which appear in an extended lemma in the corpus.

The keywords derived via the four methods are compared to a list of topics contained in the corpus. If the elicitation techniques work properly, then the key words would be indicative of the 'about-ness' of the corpus. The more the coincidences between the key words from a particular list with the projected topics, the more trustworthy the elicitation procedure via which it has been derived will be considered.

The corpus was compiled from one of the websites dedicated to Winston Churchill[1]. Churchill was chosen for this research as a well-known figure in political life. The list of topics against which the key words are tested is derived from the biography of Churchill published on the website. It is in Appendix 1. It can be expected that the speeches do not reflect every aspect from Churchill's biography that is why no complete coincidence can be expected. However, the greater the co-incidence of topics in a key word list with the biographical list of topics, the more trustworthy the procedure for deriving the key words will be considered.

The reduced frequency list is a procedure frowned upon by some for its lack of mathematical sophistication. It consists in taking the frequency list of the corpus and removing the 'function' words. As function words we treat those which are deprived of notional content – rather than those which perform grammatical functions. The outcome is also of dubious value, inasmuch as it focusses on frequency only, while the chi-square and log likelihood include a comparison with an expected value and an estimate of haphazardness. It is included for comparative purposes.

Deriving a Key word List through words with extended lemmas is done manually, via the alphabetical list produced by the Wordsmith. The words of frequency higher than 0.1 % of the entire corpus are checked for occurrence of other forms from the grammatical paradigm, or for derivatives from the same root. The concordances are then checked whether they are consistent with each other in meaning. If they are not, they are excluded from the study. As the outcome is a lengthy list, the proceeds are distilled via an index derived through the following procedure: the decimal points of the percentage of each item are multiplied by the number of members of the lemma. For example, below we see the extended lemma and derivatives of the word AIR. The first number shows how many times the word occurs in the corpus, the second – where available – presents the percentage in the corpus, used in the calculation of key words:

AERODROMES	2,00	
AEROPLANE	2,00	
AEROPLANES	6,00	
AIR	191,00	0,14
AIRBORNE	5,00	
AIRCRAFT	19,00	0,01
AIRFIELDS	3,00	

[1] http://winstonchurchill.org/resources/speeches

AIRMEN	5,00
AIRPLANES	1,00

The group contains 9 members. Two of them present a statistically significant part of the corpus: AIR 0.14 and AIRCRAFT 0.01. The sum total is 0.15. Then 15 is multiplied by 9 to give the index of 135. In this way significance is given to the relative frequency of the item and to the number of repetitions. Then the words are classified according to their extended lemma index.

A visible drawback is that some words have a shorter grammatical paradigm than others by default.

General description of the corpus

The whole corpus includes 49 discrete texts, 138 898 running words – a relatively small corpus, yet suitable for key word analysis. The cut-off point for the chi-square test was set at 0.000001 – relatively low to allow more items into the procedure.

The texts present public speeches – at election events, for the media etc., and selected parliamentary speeches.

The key word lists derived via the four different methods are presented in Table 2. For comparative purposes, they are reduced to the first 60 items

	Log likelihood	Purged frequency	Chi square	Extended lemma
N	Key word	Key word	Key word	Key word
1	OUR	GREAT	CHEERS	Great 228
2	WE	WAR	ARMORED	Government 207
3	CHEERS	BRITISH	OUR	Nation 162
4	UPON	TIME	LAUGHTER	War 155
5	WAR	WORLD	PRECIPITANCY	Britain 145
6	GREAT	GOVERNMENT	BOERS	Air plane 135
7	HAVE	CHEERS	WE	Time 120
8	WHICH	SAY	UNDERRATE	Free 105
9	LAUGHTER	UNITED	UPON	German 100
10	UNITED	COUNTRY	WAR	Power 100
11	BRITISH	PEOPLE	NAZI	Force 95
12	STATES	STATES	NAZIDOM	France 95
13	ALL	YEARS	EXPEDITIONARY	Country 92
14	OF	HOUSE	DETERRENTS	Man 88
15	ARMY	MAKE	DEFENSES	Work 88
16	HEAR	POWER	GREAT	Speak 81
17	WILL	AIR	QUARRELED	Needs 80
18	EMPIRE	RIGHT	ARMIES	People 76

19	AIR	HEAR	EMPIRE	Strength 72
20	SHALL	FAR	TARIFF	Defence 66
21	NATIONS	ARMY	EXERTIONS	Hope 64
22	US	MEN	NATIONS	World 63
23	GERMAN	THINK	WEYGAND	Fight 60
24	COUNTRY	PARTY	BOLSHEVISTS	Know 60
25	FRANCE	LONG	DEFENSE	Day 52
26	NAZI	LAST	SOCIALISTIC	Army 48
27	NATION	WELL	MILLIONS	Use 48
28	WORLD	GERMAN	UNITED	Europe 48
29	OURSELVES	FRANCE	WILLKIE	Year 46
30	MILLIONS	TRADE	SKAGERRAK	Effect 45
31	AND	EUROPE	NATION	State 44
32	MUST	FORCE	ARMY	Foundation 42
33	GOVERNMENT	LAUGHTER	TYRANNY	Friends 42
34	INDIA	LET	PEOPLES	America 40
35	ARMIES	OWN	UNMEASURED	Sea 40
36	HON	SEE	STATES	Arms 40
37	PEACE	GENERAL	OURSELVES	Lose 40
38	ENEMY	MADE	HEAR	Minister 40
39	FORCE	NEVER	MAJESTY'S	Land 36
40	POWER	FREE	WHICH	Large 36
41	NOT	FRENCH	DOMINIONS	Differ 35
42	EUROPE	HON	HAVE	Secure 35
43	TARIFF	COME	HITLER	Lead 35
44	PEOPLES	BRITAIN	ENEMY	Mean 35
45	TRADE	GOOD	BRITISH	Increase 35
46	ARE	NEW	CONANT	Number 35
47	GERMANY	THREE	INDIA	India 32
48	THAT	LIKE	GERMAN	Million 32
49	DUTY	MAN	AIR	Peace 32
50	EVERY	PEACE	SHALL	Act 30
51	HITLER	NATIONS	FRANCE	Russia 30
52	GOLD	PART	EXCHEQUER	Attack 30
53	STRENGTH	PRESENT	MANKIND	General 30
54	FIGHTING	EMPIRE	COMRADESHIP	Belief 30
55	VICTORY	TAKE	TOIL	Pass 30
56	FRENCH	COURSE	WAVELL	Battle 28
57	THE	GERMANY	UTMOST	Decide 28
58	HAS	FORCES	BRAHMINS	Island 28
59	FORCES	KNOW	COUNTRY	Ship 28
60	BE	POSITION	MEASURELESS	Organise 27

Table 2. Top 60 keywords derived via chi-square, log-likelihood, extended lemma and reduced frequency list.

It is immediately obvious that the lists differ mainly in the position of key-ness occupied by the words. A significant number of words occur in the four types of Key Word Lists. They are presented below:

CHEERS
OUR/ OURSELVES/ WE
LAUGHTER
BOERS
UPON
WAR
NAZI/NAZIDOM
DEFENSES/ DEFENSE
GREAT
ARMIES/ ARMY
EMPIRE
TARIFF
NATIONS/ NATION
MILLIONS
UNITED
TYRANNY
PEOPLES
STATES
HEAR
WHICH
HAVE
HITLER
ENEMY
BRITISH
TRADE
EUROPE
HON
PEACE
GENERAL
INDIA
GERMAN
AIR
SHALL
FRANCE
COUNTRY
MEASURELESS/UNMEASURED
STRENGTH
FIGHTING
FORCES

The small difference should be explained by the fact that the corpus is the same. This list clearly reflects topics that are typical of Churchill's career – World War 2, the British colonies, free trade, the air force, parliamentary vocabulary, as well as pronouns and connectors. The missing topics are those concerning the gold standard, the Russian threat, European arrangements after the war – more specialised and of smaller significance. It is also obvious that grammatical words – prepositions, modal verbs etc. – occur in all types of key word lists.

The words which occur exclusively in each of the key word lists are presented in Table 3.

LOG LIKELIHOOD	Chi square	Purged frequency	Extended lemma
ALL	ARMORED	SAY	Work 88
OF	PRECIPITANCY	HOUSE	Needs 80
WILL	BOERS	MAKE	Hope 64
US	UNDERRATE	RIGHT	Day 52
AND	EXPEDITIONARY	FAR	Use 48
MUST	DETERRENTS	MEN	Effect 45
NOT	QUARRELED	THINK	Foundation
ARE	WEYGAND	PARTY	42

THAT	BOLSHEVISTS	LONG	Friends 42
DUTY	SOCIALISTIC	LAST	America 40
GOLD	WILLKIE	WELL	Sea 40
VICTORY	SKAGERRAK	LET	Arms 40
THE	TYRANNY	OWN	Lose 40
HAS	STATES	SEE	Minister 40
BE	MAJESTY'S	GENERAL	Land 36
EVERY	DOMINIONS	MADE	Large 36
	CONANT	NEVER	Differ 35
	EXCHEQUER	COME	Secure 35
	MANKIND	GOOD	Lead 35
	COMRADESHIP	NEW	Mean 35
	TOIL	THREE	Increase 35
	WAVELL	LIKE	Number 35
	UTMOST	PART	Act 30
	BRAHMINS	PRESENT	Russia 30
		TAKE	Attack 30
		COURSE	Belief 30
		POSITION	Pass 30
			Battle 28
			Decide 28
			Island 28
			Ship 28
			Organise 27

Table 3. Words specific for each word list

The words in the log-likelihood key word list are predominantly function words plus the content words VICTORY, GOLD and DUTY, which signal the topics of the victory in WW2, reintroducing the gold standard, and removing duties for a range of goods.

The words in the chi squared list are items of low-frequency in the language – some have different spellings in the British and American varieties. A few personal names occur as well. In this list we can see the word DETERRENT, relating to the threat of Russia – a significant theme in Churchill's career. It may well be that Churchill introduced the idea that arming a nation can prevent others from attacking it. The words BOLSHEVISTS and SOCIALISTIC also relate to the topic of the Russian threat. TYRANNY appears to belong to the topic of the Russian influence on Eastern Europe when the respective concordance lines are consulted. It would suggest that the vocabulary of the socialist system is different from the standard corpus of the alternative political system.

The purged list contains predominantly words of general meaning. Some are related to Parliamentary practices, others – to the war, yet others are really haphazard in the range of topics. This type of list gives a very broad range of subjects related to

Churchill's career, but very few of them are genuinely typical. The overall inadequacy of this list emphasises the little significance of frequency over other factors usually considered in computing key words.

The extended lemmas list – like the purged frequency – has not been subjected to a comparison with a keyword list. That is why the list contains common words which cannot outnumber the frequency in a balanced corpus. Obviously the concern that words obtain key status because of their low frequency in a general corpus is not valid for this list. This means, however, that the indicative force of the items heavily depends on checking the respective concordances and collocates, rather than on the words in their own right. An undeniable fact is that the words do reflect highlights in Churchill's career and even though no comparisons have been made with another corpus, the list could be indicative of essential points in the corpus.

Analysis of the Key Word Lists

Scott (2015:253) notes that three types of keywords are often found: "proper nouns, keywords that human beings would recognise as key, and are indicators of the 'aboutness' of a particular text, and finally, high frequency words such as BECAUSE, SHALL or ALREADY, which may be indicators of style, rather than aboutness."

In this study we establish a taxonomy based on our results, and it is slightly different from the one proposed by Scott. The four keyword lists contain six types of entries:

- parliamentary vocabulary (despite the fact that not all the speeches were made in Parliament);
- proper names – people's names and place names;
- general substitutes;
- markers of preferred modality, syntax and deixis;
- topic indicators;
- speech mannerisms.

The tables in Appendix 2 present an analysis of the keywords in the four lists arranging them in one of the six categories. Even though our list of categories is rather broad, there are items which still remain outside of the classification. Such is the word GREAT. On the one hand it occurs together with words such as EFFORT, in which case it would belong to the category of general substitutes, on the other - it is part of the name GREAT BRITAIN, where it is definitely part of a proper name. Such nouns are marked with a question.

Where a word is marked as a topic indicator, the numbers in the respective column also show which topics are signalled by the respective key word. They correspond to those in the list of highlights for this research (Appendix 1). Most of the key words are marked to signal more than one topic, because the respective concordances reveal different occurrences related to different topics. Effectively, this happens to be the case with most of the keywords. For example, WAR combines with SOUTH AFRICAN to indicate the topic Colonial Policies, with THE GREAT to denote WWI; with EUROPEAN – for WW II.

Inasmuch as the key words are expected to give indications concerning the world view of the speaker and the about-ness of the texts, the keyword list is best suited if it contains the greatest number of words from the fifth category – called here topic indicators. The highest number of topic-indicators is contained in the extended-lemmas list – 33 out of 60, secondly – in the chi-squared list – 28 out of 60, third comes the log likelihood list – 25 out of 60. Quite expectedly, the frequency list purged of function words contains the lowest number of topic indicators – only 14 out of 60.

The proper names are very indicative of the about-ness of the texts. I find them extremely pertinent to indicate significant landmarks in the careers of the researched person. The list of people Churchill associated with cannot do without Hitler. However, it is debatable whether Weygand deserves a higher key status than, say Kitchener, or Fisher. It is difficult to assess whether the key-status is determined by the fact that the name is unusual, or by its significance for the corpus.

The general substitutes are nouns of very broad semantic properties. They often name via a combination with other words. Some of the phrases can be indicators of significant topics, like the words we called 'topic indicators'. That is why they reinforce the need to use key phrases rather than single key words. However, some combinations then may not live up to the key status.

The speech mannerisms are different from the famous catch phrases known for Churchill. Neither IRON, nor CURTAIN has a key status according to any of the classifications, despite the fact that 5 occurrences of the phrase are available in the corpus. At the same time, EFFORT is a key word and in combination with WAR. Together with synonymous phrases, such as PRODIGIOUS, NATION-WIDE, SUPREME etc., this appears a phrase widely used by Churchill.

This is where a water tight borderline is needed between cultural and statistically established key words. While IRON CURTAIN is a cultural key expression for Churchill, known and popularised as a land mark of his speech, a scrupulous statistical analysis never draws any attention to it. Instead, such an analysis claims that Churchill persistently referred to WAR EFFORT. Although IRON CURTAIN never achieved statistical significance, the phrase had an undoubted impact on society by virtue of its uniqueness, though not by a frequent use.

But the key words need not only relate to topics in Churchill's career. As can be seen – and this can be no surprise – not a word suggests about Churchill's terms as prisoner of war, or of his love for polo. This may be due to the selection made by the web site constructors. The availability of Parliamentary vocabulary, in its part, is indicative of Churchill's operation in parliament and cannot be overlooked when portraying him.

Conclusions

The key word lists included in this research are indeed indicative of highlights in Churchill's career. The most indicative is the list of extended lemmas and the least – the reduced frequency list.

The log-likelihood, although it is widely preferable for specialists, appears – on this occasion – too cluttered with function words and general substitutes. In view of having more notion words of specific meaning, evocative of topics, the chi-square leads to a greater number of indicative words.

The most evocative key word list is the extended lemma list. Linguistic software, such as Wordsmith, however, does not derive such a statistic. It may also be difficult to derive automatically, inasmuch as the decision which parts of the lemma need to be included, and which derivative words may need human involvement. Certainly, the option to merge entries is very helpful in the matter.

The research leads to the conclusions that the list of key words which projects items appearing in an extended lemma in a corpus indeed is indicative of at least as many topics as the typically derived chi square and log likelihood. More work needs to be done on the procedures for deriving it.

References

Baker, P. (2004). Querying keywords: questions of difference, frequency and sense in keywords analysis" *Journal of English Linguistics*, *32*(4), 346-359.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Davies, Mark. (2004). BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61-74.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2005). *Lexical Priming: A new theory of words and language*. Routledge.

Kilgarriff, A. (1996). *Which Words are Particularly Characteristic of Text? A Survey of Statistical Approaches.* Information Technology Research Institute, University of Brighton. Retrieved from https://www.kilgarriff.co.uk/Publications/1996-K-AISB.pdf

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Phillips, M. (1989). *Lexical Structure of Text Discourse* Analysis Monograph No. 12, Birmingham: English Language Research, University of Birmingham.

Scott, M. (1997). PC Analysis of Key words - and Key Key Words. *System, 25*(2), 233-245.

Scott, M (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In Mohsen Ghadessy, Alex Henry, Robert L. Roseberry (Eds.), *Small corpus studies and ELT: theory and practice*. John Benjamins B.V.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails…. In Bondi, M. & Scott, M. (Eds.), *Keyness in texts* (pp. 43–58). John Benjamins B.V. doi:10.1075/scl.41

Scott, M. (2012). *WordSmith Tools* version 6 [Computer Software], Stroud: Lexical Analysis Software. Retrieved from http://www.lexically.net/wordsmith/index.html

Scott, M. (2015). *Wordsmith Tools Manual*. Lexical Analysis Software Ltd. Retrieved from http://lexically.net/downloads/version6/HTML/index.html?getting_started.htm

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins B.V.

Sinclair, J. (1996). The Search for Units of Meaning. *Textus IX*. 75-106.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. London: Blackwell.

Stubbs, M. (2010). Three concepts of keywords. In Bondi, M. & Scott, M. (Eds.), *Keyness in texts* (pp. 21–42). John Benjamins B.V. doi:10.1075/scl.41

Tarasheva, E. (2011). *Repetitions of Word Forms in Texts*. Cambridge Scholars Publishing.

Utka, A. (2004). Analysis of George Orwell's novel *1984* by statistical methods of corpus linguistics.' (Bachelor's thesis, Kaunas Vytautas Magnus University, Kaunas, Lithuania). Retrieved from http://donelaitis.vdu.lt/publikacijos/adrtmain.htm

Williams, R. (1976/1983). *Keywords: A Vocabulary of Culture and Society*. London: Fontana Press.

Appendix 1

1. Army service
2. War correspondent
3. Polo-player
4. Freemason
5. Prisoner Of War
6. Proponent of Free trade
7. Colonial Policy Supporter
8. Navy Reform Proponent
9. Airplane Warfare Proponent
10. Labour legislation
11. Mental Deficiency Act 1913
12. The Russian threat
13. Irish Independence
14. Suffragettes
15. Handling strikes
16. Returning the golden standard
17. Anti-fascist action
18. Anti-abdication
19. Co-operation with America
20. Alliance with France
21. Engineering the Yalta agreement
22. Partisan of United States of Europe, sponsored by USA & UK

Appendix 2

The Chi-square list analysed

	Chi-square clalculation				Topics covered
N	Key word	Freq.	%	Texts	
1	CHEERS	251	0.18	699	Parliamentary vocab
2	ARMORED	14	0.01	6	17
3	OUR	1,007	0.73	93,455	Preferred deixis
4	LAUGHTER	135	0.10	2,068	Parliamentary vocab
5	PRECIPITANCY	10	2		Mannerism
6	BOERS	13	13		7, 1
7	WE	1,724	1.24	300,833	Preferred deixis
8	UNDERRATE	13	16		12, 17
9	UPON	384	0.28	22,806	Mannerism
10	WAR	408	0.29	27,222	17
11	NAZI	61	0.04	754	17
12	NAZIDOM	5	0		17
13	EXPEDITIONARY	17	0.01	57	17, 8, 9
14	DETERRENTS	14	0.01	37	22, 12
15	DEFENSES	5	1		17, 12
16	GREAT	447	0.32	46,647	?
17	QUARRELED	4	0		Mannerism
18	ARMIES	57	0.04	998	17, 1, 12
19	EMPIRE	106	0.08	3,503	7
20	TARIFF	45	0.03	666	6
21	EXERTIONS	17	0.01	87	Mannerism
22	NATIONS	109	0.08	4,115	7,17, 12
23	WEYGAND	4	1		Proper name
24	BOLSHEVISTS	4	1		12
25	DEFENSE	24	0.02	203	17, 22, 12
26	SOCIALISTIC	7	12		12, 21
27	MILLIONS	80	0.06	2,638	Mannerism
28	UNITED	228	0.16	19,030	22
29	WILLKIE	3	0		Propername
30	SKAGERRAK	3	0		Placename
31	NATION	92	0.07	3,567	General substitute
32	ARMY	162	0.12	10,862	1, 17, 8, 9
33	TYRANNY	25	0.02	278	12, 17
34	PEOPLES	56	0.04	1,503	General substitute
35	UNMEASURED	7	16		Mannersim of speech
36	STATES	207	0.15	17,873	General substitute
37	OURSELVES	96	0.07	4,432	Preferred Deixis
38	HEAR	172	0.12	13,177	Parliamentary vocab
39	MAJESTY'S	32	0.02	535	Parliamentary vocab
40	WHICH	1,289	0.93	366,196	Syntactic Preferencs
41	DOMINIONS	18	0.01	164	7
42	HAVE	1,477	1.06	448,684	Modus
43	HITLER	46	0.03	1,171	17
44	ENEMY	75	0.05	3,057	17
45	BRITISH	287	0.21	35,530	Nationality name
46	CONANT	3	1		Proper name

47	INDIA	89	0.06	4,295	7
48	GERMAN	146	0.11	10,870	17
49	AIR	191	0.14	18,415	9
50	SHALL	197	0.14	19,817	Preferred modality
51	FRANCE	145	0.10	11,552	20, 22
52	EXCHEQUER	36	0.03	825	6, 16, 15
53	MANKIND	34	0.02	738	General substitute
54	COMRADESHIP	11	71		Mannerism
55	TOIL	16	0.01	176	Mannerism
56	WAVELL	5	12		Proper name
57	UTMOST	26	0.02	504	Mannerism
58	BRAHMINS	6	20		7
59	COUNTRY	218	0.16	27,959	General substitute
60	MEASURELESS	5	13		Mannerism

The loglikelihood

	Log likelihood				Topics covered
N	Key word	Freq.	%	RC. Freq.	
1	OUR	1,007	0.72	93,455	Preferred deixis
2	WE	1,724	1.24	300,833	Preferred deixis
3	CHEERS	251	0.18	699	Parliamentary vocab
4	UPON	384	0.28	22,806	Mannerism
5	WAR	408	0.29	27,222	6, 7, 8, 9 , 17
6	GREAT	447	0.32	46,647	?
7	HAVE	1,477	1.06	448,684	Preferred modality
8	WHICH	1,289	0.93	366,196	Preferred syntax
9	LAUGHTER	135	0.10	2,068	Parliamentary vocab
10	UNITED	228	0.16	19,030	19
11	BRITISH	287	0.21	35,530	Proper name
12	STATES	207	0.15	17,873	19
13	ALL	899	0.65	277,566	?
14	OF	5,755	4.14	3,049,564	?
15	ARMY	162	0.12	10,862	7, 8, 9, 17
16	HEAR	172	0.12	13,177	Parliamentary vocab
17	WILL	816	0.59	251,179	Preferred modality
18	EMPIRE	106	0.08	3,503	7
19	AIR	191	0.14	18,415	9
20	SHALL	197	0.14	19,817	Preferred modality
21	NATIONS	109	0.08	4,115	General substitute
22	US	388	0.28	80,226	Preferred deixis
23	GERMAN	146	0.11	10,870	17
24	COUNTRY	218	0.16	27,959	General substitute
25	FRANCE	145	0.10	11,552	17, 20, 22
26	NAZI	61	0.04	754	17
27	NATION	92	0.07	3,567	General substitute
28	WORLD	287	0.21	53,806	General substitute
29	OURSELVES	96	0.07	4,432	Preferred deixis
30	MILLIONS	80	0.06	2,638	?
31	AND	4,808	3.46	2,624,341	Preferred syntax
32	MUST	324	0.23	69,099	Preferred modality
33	GOVERNMENT	285	0.21	56,343	Parliamentary vocab
34	INDIA	89	0.06	4,295	6

35	ARMIES	57	0.04	998	1, 7, 8, 9, 17
36	HON	121	0.09	10,692	Parliamentary vocab
37	PEACE	111	0.08	8,707	7, 17, 22
38	ENEMY	75	0.05	3,057	7, 8, 9, 17
39	FORCE	140	0.10	15,475	8, 9
40	POWER	197	0.14	31,627	General substitute
41	NOT	1,052	0.76	431,075	Preferred modality
42	EUROPE	141	0.10	16,908	17, 20, 22
43	TARIFF	45	0.03	666	6
44	PEOPLES	56	0.04	1,503	7
45	TRADE	145	0.10	19,818	6
46	ARE	1,070	0.77	458,368	Modality
47	GERMANY	101	0.07	9,399	17
48	THAT	2,090	1.50	1,052,259	Syntax
49	DUTY	93	0.07	7,869	6, 17
50	EVERY	201	0.14	39,156	?
51	HITLER	46	0.03	1,171	17
52	GOLD	86	0.06	7,574	16
53	STRENGTH	83	0.06	6,957	General substitute
54	FIGHTING	75	0.05	5,528	8, 9, 17, 20
55	VICTORY	75	0.05	5,547	6, 17
56	FRENCH	122	0.09	16,879	20, 22
57	THE	9,754	7.02	6,055,105	?
58	HAS	648	0.47	252,703	?
59	FORCES	100	0.07	11,656	?
60	BE	1,356	0.98	651,535	

The reduced frequency list

	Purged frequency				Topics covered
N	Word	Freq.	%	Texts	
38	GREAT	447	0.32	46	?
41	WAR	408	0.29	39	1, 17
66	BRITISH	287	0.21	43	Place name
67	TIME	287	0.21	44	General substitute
68	WORLD	287	0.21	46	General substitute
69	GOVERNMENT	285	0.21	31	Parliamentary vocab
72	CHEERS	251	0.18	10	Parliamentary vocab
75	SAY	229	0.16	43	General substitute
77	UNITED	228	0.16	41	19
79	COUNTRY	218	0.16	36	General substitute
81	PEOPLE	208	0.15	39	General substitute
82	STATES	207	0.15	40	General substitute
83	YEARS	205	0.15	42	General substitute
85	HOUSE	200	0.14	30	Parliamentary vocab
86	MAKE	198	0.14	42	General substitute
87	POWER	197	0.14	40	?
89	AIR	191	0.14	26	9, 17
94	RIGHT	173	0.12	41	Parliamentary vocab
95	HEAR	172	0.12	13	Parliamentary vocab
96	FAR	168	0.12	38	?
98	ARMY	162	0.12	22	1, 17

99	MEN	162	0.12	40	General substitute
101	THINK	161	0.12	33	General substitute
104	PARTY	152	0.11	31	Parliamentary vocab
107	LONG	149	0.11	43	?
108	LAST	148	0.11	42	?
109	WELL	148	0.11	41	?
110	GERMAN	146	0.11	27	17
112	FRANCE	145	0.10	29	20
113	TRADE	145	0.10	18	6
119	EUROPE	141	0.10	29	22, 17
121	FORCE	140	0.10	31	8,9
123	LAUGHTER	135	0.10	13	Parliamentary vocab
124	LET	135	0.10	38	?
125	OWN	135	0.10	42	?
127	SEE	134	0.10	34	?
130	GENERAL	132	0.10	31	?
131	MADE	131	0.09	37	?
132	NEVER	131	0.09	41	?
134	FREE	130	0.09	32	6
140	FRENCH	122	0.09	25	20, 22
141	HON	121	0.09	10	Parliamentary vocab
142	COME	120	0.09	38	?
144	BRITAIN	116	0.08	36	Place name

The extended lemma

	Extended lemmas	Topics covered
N	Key word	
1	Great 228	?
2	Government 207	Parliamentary vocab
3	Nation 162	17, 6, 22, 20, 19
4	War 155	17
5	Britain 145	Place name
6	Air plane 135	9
7	Time 120	General substitute
8	Free 105	17, 6
9	German 100	17
10	Power 100	8,9, 17
11	Force 95	8, 9, 17
12	France 95	20, 11
13	Country 92	General substitute
14	Man 88	General substitute
15	Work 88	General substitute
16	Speak 81	Parliamentary vocab
17	Needs 80	General substitute
18	People 76	General substitute
19	Strength 72	8, 9, 17
20	Defence 66	8,9, 17, 20, 19,22
21	Hope 64	General substitute
22	World 63	General substitute
23	Fight 60	7,8, 17
24	Know 60	General substitute
25	Day 52	General substitute

26	Army 48	8, 9, 17
27	Use 48	General substitute
28	Europe 48	17, 20, 22, 19
29	Year 46	General substitute
30	Effect 45	General substitute
31	State 44	General substitute
32	Foundation 42	6,7, 19
33	Friends 42	Parliamentary vocab
34	America 40	22
35	Sea 40	8, 17
36	Arms 40	8,9, 17
37	Lose 40	6, 17, 20
38	Minister 40	Parliamentary vocab
39	Land 36	17
40	Large 36	General substitute
41	Differ 35	General substitute
42	Secure 35	17, 21, 6, 22
43	Lead 35	Parliamentary vocab
44	Mean 35	General substitute
45	Increase 35	6, 10, 8, 9, 22
46	Number 35	8,9, 17
47	India 32	7
48	Million 32	General substitute
49	Peace 32	17, 8,9
50	Act 30	17, 20, 19
51	Russia 30	17, 22, 12
52	Attack 30	17, 12, 19
53	General 30	15, 17, 20, 19
54	Belief 30	6, 7, 21, 22
55	Pass 30	Parliamentary vocab
56	Battle 28	17, 20
57	Decide 28	6, 15
58	Island 28	17, 7
59	Ship 28	17, 8, 19
60	Organise 27	19, 21, 22