Calculating Key Words In Texts: Methods And Relevance

Elena Tarasheva, Ph.D.

Abstract: *In this paper the key word lists for a corpus derived via different mathematical procedures are compared to a summary of significant points in the corpus to check whether they faithfully reveal the highlights, as corpora analyses purport to claim. A new statistical procedure is put forward, as yielding the greatest number of coincidences with the summary.*

Key words: *corpus linguistics, corpus assisted discourse studies, chi square, log likelihood.*

Заглавие*: Извличане на ключови думи от текстове: методи и пригодност*
Автор: *Елена Тарашева*
Резюме: *В този доклад се представя изследване на списъци от ключови думи, съставени с помощта на различни математически процедури от експериментален корпус. Всеки от списъците се съпоставя с кратко обобщение на важните точки за корпуса, за да се прецени къде има най-голям брой съвпадения и дали действително списъците с ключви думи могат да ориентират относно съдържанието на корпуса, както се твърди в Корпусната лингвистика. Предлага се нова статистическа процедура, която демонстрира най-голямо съвпадение с обобщението.*

INTRODUCTION

In the paper four different methods for establishing key words in texts are compared with a view of the results that they yield. Certain words are considered KEY to a corpus and they are believed to be revealing of the 'about-ness' of the corpus. The branch of linguistics Corpus assisted discourse studies depends heavily on such claims. Different statistical procedures for calculating which words are KEY, however, lead to different components of the key word lists. This erodes the trustworthiness of the concept and an attempt is made here to justify choices. This research explores what is it that the key words from different calculations lead to on a corpus of speeches by Winston Churchill by comparing the list to a list of highlights in Churchill's career.

THE CONCEPT OF KEYWORDS

Scott and Tribble [1] base their approach to establishing key words on repetitive reference. If a proposition – as suggested by Kintsch and van Dijk [2] – or a sentence – as suggested by Hoey [3] – is referred to repetitively, then it should have more importance about the text as a whole. A significant distinction to make here is that propositions do not necessarily contain the same lexical form. Repeating a proposition can be done using a synonym, a derivative or a paraphrase, all of which would not include the same word repeated.

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.

The issues that exist with procedures for deriving key words have been discussed many times and they include the fact that reference often takes place without a direct lexical repetition. Even most style guides warn against repeating words and phrases. Research [4] has shown that a Prime Minister announces reshuffles in his cabinet without a single mention of the words RESHUFFLE or CHANGE. Thus, a researcher trying to get to the 'about-ness' of the speech is unlikely to do so via a list based on the frequency of lexical repetitions.

Several methods of deriving key words exist. The creator of one of the most popular software products for linguistic analysis Wordsmith, Scott [5] describes deriving keywords in the following way:

"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". The notion underlying this is therefore "outstandingness" based on comparison. In this tool (Key words – E.T.), as in Word List, a number of detailed statistics are made available, but the chief interest of the tool lies in its ability to get at text "aboutness"".

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 [6]. 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 expected – 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:

$$G^2=2\sum x_y (\log_4 x_4 - \log_e m_e)$$

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 [6].

Kilgarriff [7], having compared the chi-square and log-likelihood (also known as G-square) measures, preferred the G-square. Dunning [8] 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.

METHOD AND PROCEDURE

For the purposes of this study a corpus was compiled from one of the websites dedicated to Winston Churchill [9]. Churchill was chosen for this research as a well-known figure in political life. Therefore, it would be visible which aspects of his life are reflected in a key word list. However, to make matters more precise, based on his biography, a list of landmark events in his life was derived against which each key word list was tested.

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

The software used for the research is Wordsmith [10]. The reference corpus in all the cases was the British National Corpus, as the most neutral of existing options.

Four types of key word lists were derived from the corpus:
1. The typical chi-squared list derived automatically via the software Wordsmith tools;
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 (called here REDUCED FREQUENCY LIST);
4. The list of words which appear in an extended lemma in the corpus (called here THE EXTENDED LEMMA LIST).

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.

The fourth type of analysis 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 by Tarasheva [11] reveals that concepts central to research articles occur in different forms because they are discussed in different types of reference - generic, specific, classificatory etc., thereby – in different forms of the respective lemma. The same holds true for items central to short story narratives and political speeches.

In English word lemmas are restricted, but are still indicative. Apart from forms including grammatical markers, semantic derivatives can also be seen as part of an extended lemma, as, for instance, PLANE and AIRPLANE.

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: in figure 1. we see the extended lemma of the word AIRPLANE:

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	
AIRMEN	5,00	
AIRPLANES	1,00	

Figure 1. The extended lemma of AIR

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.

The keywords derived via the four methods are compared to the list of topics significant for Churchill's life compiled for this research. The comparison of the key lists to the themes in Churchill's life is intended to reveal whether a researcher is likely to learn about Churchill through the key word lists – a claim inherent in the efforts of many discourse analysts. The expectation is not that every single aspect of Churchill's life should be reflected in the keywords for his speeches. However, the list should be indicative of a fair amount, because the speeches were selected as representative of Churchill's career.

DATA DESCRIPTION

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.

First, we take a look at the key word lists derived via the four different methods. They are presented in Table 1. For comparative purposes, they are reduced to the first 10 items.

	Log likelihood	Reduced 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

Table 1. The Key Word Lists Juxtaposed – the first 10 items

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.

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.

The words which occur exclusively in each of the lists are presented in Table 2:

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 42
ARE	WEYGAND	PARTY	Friends 42
THAT	BOLSHEVISTS	LONG	America 40
DUTY	SOCIALISTIC	LAST	Sea 40
GOLD	WILLKIE	WELL	Arms 40
VICTORY	SKAGERRAK	LET	Lose 40
THE	TYRANNY	OWN	Minister 40
HAS	STATES	SEE	Land 36
BE	MAJESTY'S	GENERAL	Large 36
EVERY	DOMINIONS	MADE	Differ 35

Table 2. Unique Words in Key Word Lists

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 reduced frequency list contains predominantly words of general meaning. Some are related to Parliamentary practices, others – to the war, yet others are really haphazard. 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 do not 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 DATA

Scott [12] 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 below 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. 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. To avoid this type of ambiguity, it might make sense to elicit phrases rather than single words, as has been suggested by other researchers.

Space restrictions prevents us here from presenting the entire Key Word Lists. The tables are indicative of the procedure, but they do not give the whole picture.

	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

Table 3.The Key Word List Derived via Chi Square - analysis

	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

Table 4.The Key Word List Derived via Log Likelihood - analysis

	Reduced 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

Table 5.The Key Word List Derived via Reduced Frequency - analysis

	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

Table 6.The Key Word List Derived via Extended Lemmas – analysis

Inasmuch as the discourse is 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 reduced 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 used here, 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 speeches, a scrupulous statistical analysis never draws any attention to it. Instead, such an analysis claims that Churchill persistently referred to WAR EFFORT – and this is the truth of it. 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 – who ignored speeches on these topics. The availability of Parliamentary vocabulary, in its part, is indicative of Churchill's operation in parliament and cannot be overlooked when portraying him.

CONCLUSION

1. The key word lists included in this research are indeed indicative of the major highlights in Churchill's career. The most indicative is the list of extended lemmas and the least – the reduced frequency list.
2. Each key word relates to more than one topic area as formulated for this research.
3. The log-likelihood, although it is widely prefered by 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.
4. 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.

REFERENCES

[1] Scott, M. & Tribble, C. Textual Patterns: Key words and corpus analysis in language. John Benjamins B.V. 2006

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

[3] Hoey, M. Patterns of Lexis in Text. Oxford: OUP. 1991

[4] Tarasheva, E. “Езиков и реторичен анализ на две политически речи” в: Янкова, Д. (ред.) Език, Литература, Култура. Юбилеен сборник на Департамент “Приложна Лингвистика”. Нов Български Университет стр. 26-36. 2004.

[5] Scott, M. “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 (ed.) Small corpus studies and ELT : theory and practice John Benjamins B.V. 2001.

[6] Oakes, M. Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press. 1998.

[7] Kilgarriff, A. 'Which Words are Particularly Characteristic of Text? A Survey of Statistical Approaches', Information Technology Research Institute, University of Brighton, 6 March. 1996.

[8] Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, Volume 19, number 1: 61-74. 1993.

[9] Churchill website: http://winstonchurchill.org/resources/speeches

[10] Scott, M. WordSmith Tools version 6, Stroud: Lexical Analysis Software. 2012.

[11] Tarasheva, E. Repetitions of Word Forms in Texts. Cambridge Scholars Publishing. 2011

[12] Scott, M. Wordsmith Tools Manual. Lexical Analysis Software Ltd. 2015.

Contact details:
Associate Professor Elena Tarasheva, Ph.D. Department of English Studies, New Bulgarian University, e-mail: etarasheva@nbu.bg