‘Keywords Method’ versus ‘Calcul des Spécificités’

A comparison of tools and methods

Ann Bertels and Dirk Speelman
KU Leuven

This paper explores two tools and methods for keyword extraction. As several tools are available, it makes a comparison of two widely used tools, namely *Lexico3* (Lamalle et al. 2003) and *WordSmith Tools* (Scott 2013). It shows the importance of keywords and discusses recent studies involving keyword extraction. Since no previous study has attempted to compare two different tools, used by different language communities and which use different methodologies to extract keywords, this paper aims at filling the gap by comparing not only the tools and their practical use, but also the underlying methodologies and statistics. By means of a comparative study on a small test corpus, this paper shows major similarities and differences between the tools. The similarities mainly concern the most typical keywords, whereas the differences concern the total number of significant keywords extracted, the granularity of both probability value and typicality coefficient and the type of the reference corpus.

Keywords: keyword extraction, probability value, typicality coefficient, reference corpus

1. Introduction

The aim of this paper is to compare two tools and methods for keyword extraction.[1] When generated by a computer, ‘keywords’ are those words that occur significantly more often in a specific corpus than one would predict or expect on the basis of their frequency in a (usually large and general) reference corpus (Scott 2001; Scott & Tribble 2006). To identify the keywords, frequencies in the specific corpus are compared to frequencies in the reference corpus, with respect to the total number of words in both corpora. In other words, the observed frequency in the specific corpus is compared to the expected frequency, which

International Journal of Corpus Linguistics 18:4 (2013), **536–560.** DOI 10.1075/ijcl.18.4.04ber
ISSN 1384–6655 / E-ISSN 1569–9811

is determined on the basis of observations in the reference corpus. If there is a difference between the observed frequency and the expected frequency and if this difference is statistically significant, it allows us to identify the keywords in the specific corpus. Since no previous study attempted to compare two different tools for keyword extraction, this paper seeks to fill the gap. At the same time, it seeks to bridge two language communities, by comparing *WordSmith Tools* and *Lexico3*, which are predominantly used in the English and French speaking community respectively.

Recent research involving identification and extraction of keywords in specialised corpora has been carried out with the help of several tools. Berber-Sardinha (1999a, 1999b) uses *WordSmith Tools* (Scott 2013) which includes the Keywords tool to compare a small corpus of business reports (3,355 tokens) to a reference corpus of 17 reports (95,541 tokens). Vangehuchten (2004) uses the same tool to examine Spanish for specific purposes. She identifies core vocabulary by comparing a small corpus of specialised texts of about 120,000 tokens to a general language reference corpus of about 19.4 million tokens. *WordSmith Tools* has already been used for French (Fonseca-Greber 2007). A very early keywords method is described in Lyne (1985), also for French.

Another tool is *Lexico3* (Lamalle et al. 2003), used for instance by Zimina (2004) for keyword extraction in parallel corpora (French-English). Drouin (2003, 2004) uses the underlying methodology for terminology extraction in a technical corpus, but designed a new tool, *TermoStat* (Drouin 2003), which merges the two corpora into one global heterogeneous corpus. This merger is unusual in that it detects specific technical vocabulary using a technical specialised corpus and a non-technical reference corpus. Lemay et al. (2005) conduct a comparative study, evaluating two methods for extraction of single word terms in a specialised IT corpus. Both methods are based on 'Calculation of Specificities', a method implemented in the *TermoStat* program (Drouin 2003, 2004). The two methods differ, however, with respect to the reference corpus. The first method compares the specialised IT corpus (600,000 tokens) to a general language reference corpus (30 million tokens). The second method compares each of the six topical sub-corpora of the specialised corpus to the entire specialised corpus and joins the six keyword lists. Both methods are evaluated by comparing their keyword list to the content of two specialised dictionaries, and are. both useful for extraction of single word terms. Beyond studies on language for specific purposes, keywords are also often used in the area of corpus-assisted discourse studies (Baker 2004; Bondi & Scott 2010).

Besides *Lexico3* and *WordSmith Tools*, other tools are available for keyword extraction, for example *AntConc* (Anthony 2011), *Wmatrix* (Rayson 2009) and the *AV Frequency List Tool* (Speelman 1997). Most previous studies use and describe

one tool and methodology for keyword extraction. Lemay et al. (2005) conduct a comparative study on keyword extraction and limit their experiments to one underlying methodology (Calculation of Specificities). The underlying methods log-likelihood ratio and Fisher's Exact have already been discussed together previously for collocation extraction (Moore 2004; Evert 2007). Baron et al. (2009) carries out a survey of keyness methods, including the chi-squared test (χ^2), the log-likelihood ratio test statistic and the Fisher's Exact test. Fisher's Exact may be used for tables with small expected frequencies, as an alternative to the chi-squared test (Baron et al. 2009).

The main questions to be addressed in this paper are the following:

i. How can keywords be extracted? What is the methodology behind keyword extraction?
ii. Which tool or method is most appropriate for a given research purpose and a given research corpus?

In this paper, the research purpose that we use to compare tools and methods is the search for a fine-grained typicality ranking of the keywords of a technical test corpus with respect to a general language reference corpus.

We first describe the tool *Lexico3* and the underlying method Calculation of Specificities (Section 2), as well as *WordSmith Tools* and the underlying Keywords Method (Section 3). Next, a comparative study is carried out to identify major similarities and differences between the tools in terms of probability value, typicality coefficient and reference corpus (Section 4). Finally, this paper presents some conclusions and suggestions for further research (Section 5).

2. *Lexico3* and the underlying method Calculation of Specificities

Lexico3 (or L3) was developed by SYLED-CLA2T (Paris 3) and contains several tools for lexicometric analysis and textual statistics, such as the identification of concordances and co-occurrences, basic descriptive statistics (counts and histograms) and different types of multivariate textual data analysis. Lamalle & Salem (2002) illustrate how Generalized Types or *Tgen(s)* can be identified and Sansonetti (2003) uses L3 for cluster analysis. L3's tool for identifying keywords (or "spécificités") is based on the underlying methodology called Calculation of Specificities (or "calcul des spécificités") (Lafon 1984) (see Section 2.2). Even though L3 is typically used for other (more lexicometric) research applications, it provides an alternative to *WordSmith Tools* for keyword identification and extraction.

2.1 The tool L3

L3 has been designed to identify typical vocabulary within one section of a complete corpus. The purpose is to determine whether a word in the selected section occurs more or less frequently than would be expected by chance or with respect to the complete corpus. The input is textual data with different sections. In most cases, the size of the analysed section *t* is 1/10 of the size of the entire corpus *T*. Every section should be delimited by a proposed delimiter, e.g. $. Since L3 only allows part-whole comparison, our technical test corpus and the general language reference corpus are incorporated into one large hybrid corpus. A delimiter $ precedes the specialised corpus similarly to the other parts of the reference corpus. However, in order to identify French keywords, the delimiters – and / have to be deleted; otherwise compounds with hyphen and slash such as *machine-outil* and *m/mn* are lost, since they would be considered as two separate words. One or more sections can be selected for keyword calculation and extraction. The default probability value in L3 is 5 (or $p < 0.05$), but both minimal frequency and *p*-value can be modified.

A probability value (or *p*-value) checks whether the null hypothesis can be rejected. According to the null hypothesis, there is no difference between observed and expected frequency, or no difference between the frequencies in the specific and the reference corpus, with respect to the total number of words in both corpora. A *p*-value smaller than 0.05 is generally considered as a cut-off point for safely rejecting the null hypothesis, in which case differences are statistically significant and the likelihood that they are due to chance is only 5%. At $p < 0.1$, 10% of the differences are merely due to chance. At $p < 0.001$, differences are highly significant with an error rate of 0.1%.

After calculating all the keywords or "spécificités", the left side of the screen in L3 shows four columns (see Table 1): (i) all the positive keywords, typical of the section (or technical test corpus), (ii) total frequency throughout the corpus, (iii) frequency in the section, and (iv) typicality coefficient. Note that positive keywords are those words which are unusually frequent in the target section in comparison with the reference corpus. A higher coefficient value indicates a lower probability for the word to appear as often in the target section as it appears throughout the corpus. The results can be saved in a report for further analysis. According to the manual, this typicality coefficient x is an exponent (10^{-x}) and it indicates the degree of significance of the deviation. A typicality coefficient value of 2 stands for a probability value of 10^{-2} ($p \leq 0.01$), which means that the *p*-value of keywords with typicality coefficient 2 ranges between 0.001 and 0.01 ($0.001 < p \leq 0.01$). The same goes for a coefficient value of 1 ($0.01 < p \leq 0.1$). To ensure all keywords in the L3 list are statistically significant ($p < 0.05$), all keywords with a coefficient

value of 1 must be eliminated, even if some of them respect the rule of $p < 0.05$. Unfortunately, the same (coarse-grained) coefficient value is attributed to several keywords at a time (see Section 4.2.2 for concrete data comparison). Besides, the most typical keywords all have the same typicality indication three asterisks (***, see Table 1), which impedes a clear-cut typicality ranking.

Table 1. Output L3: specificities and frequency information in 4 columns

Forme	Frq. tot.	Frq. partie	Coeff.
diamètre	156	154	***
course	156	112	***
moteur	150	119	***
gamme	154	115	***
constructeur	161	107	***
0	221	120	***
précision	230	203	***
rapide	178	116	***
jusque	195	168	***
déplacement	125	90	***
conception	131	105	***
surface	119	84	***
commande	124	84	***
acier	131	126	***
tournage	146	135	***
plaquette	147	146	***
fabrication	133	104	***
Trametal	140	140	***
broche	256	256	***
système	502	248	***

2.2 The underlying method: Calculation of Specificities

The earliest methodological approach for identifying keywords, called 'Calculation of Specificities' ("Calcul des Spécificités") (Lafon 1984; Müller 1992) and implemented in L3, calculates frequency variations and variability in a corpus divided into various sections (Labbé & Labbé 2001). A comparison of the frequency of a word in the extracted section to the frequency of the same word in the whole corpus indicates whether the word appears more frequently in the section than predicted, through the underlying statistical analysis established via the Fisher's Exact test statistic, based on the exact probabilities of hypergeometric distribution (see Appendix 1).

2.2.1 *Hypergeometric distribution in a text corpus*

As discussed in detail in Appendix 1, a hypergeometric distribution is a discrete probability distribution. It describes the number of successes in a sequence of n draws (fixed size sample) from a discrete and finite population, without replacement. For example, in a population of N balls, m balls are red and the others ($N–m$) are white. If we select and withdraw n balls from the population, what is the probability that exactly k selected balls are red? Or, to be more precise, in a box with 20 balls, 14 balls are red and 6 balls are white. If 8 balls are randomly extracted, the likelihood that exactly 6 selected balls are red is described by a hypergeometric distribution.

The selection of one section of a text corpus can be considered as a sample of words taken out of a population of words. A selected section does not include the same paragraph twice. For the analysis of words, the most important quantitative data are the total number of words in the corpus and the frequency of these words. After taking out the first word, the total number of words in the text corpus changes. Therefore, a hypergeometric distribution, characterising a sample drawn from a population without replacement, seems appropriate for the analysis of a section of a text corpus. When one section is analysed in comparison with the whole text corpus, the general formula (see Figure 5 in Appendix 1) can be adapted to textual data. Of interest here is not the probability of having exactly 6 red balls (see Appendix 1 for more details), but the probability of having exactly the same relative frequency in the sample section compared to the population in the text corpus, in other words the probability that the observed frequency equals the expected frequency, the latter being the virtual frequency in the section based on the observed frequency in the whole corpus. As a result, the Calculation of Specificities indicates whether a word's frequency in a section is normal, in that its observed frequency equals the expected frequency. When the observed frequency of a word largely exceeds the expected frequency, the calculated probability is infinitely small. The formula used by Lafon (1984: 57) (see Figure 1) calculates the probability that a word with frequency f throughout the corpus (size T) occurs k times in the section or sample i (size t_i), according to the hypothesis of equiprobability of the samples and hypergeometric distribution.

$$\text{Prob}\,(X=k) = \frac{\binom{f}{k}\binom{T-f}{t_i-k}}{\binom{T}{t_i}}$$

Figure 1. Formula for hypergeometric distribution in a text corpus

2.2.2 *Calculating probability in a large text corpus*

Since we are dealing with text corpora of several thousand or even millions of words, the factorials of the definition given by Lafon (1984: 57) and detailed below (see Figure 2), lead to astronomical numbers that are hard to process. Therefore, Lafon (1984: 65–66) suggests using logarithms via the following formula (see Figure 3). The Fisher's Exact test statistic used in the hypergeometric distribution is computationally expensive and a sophisticated implementation is necessary to avoid numerical instabilities (Evert 2007). However, it is very useful for low-frequency data (Evert 2007).

$$\begin{aligned}\text{Prob}\,(X=k) &= \frac{\binom{f}{k}\binom{T-f}{t_i-k}}{\binom{T}{t_i}} \\ &= \frac{f!}{(f-k)!k!}\cdot\frac{(T-f)!}{(T-f-t_i+k)!(t_i-k)!}\cdot\frac{(T-t_i)!t_i!}{T!} \\ &= \frac{f!(T-f)!(T-t_i)!t_i!}{T!k!(f-k)!(t_i-k)!(T-f-t_i+k)!}\end{aligned}$$

Figure 2. Detailed formula for hypergeometric distribution in a text corpus

$$\begin{aligned}\log \text{Prob}\,(X=k) = {} & \log f! + \log (T-f)! + \log t_i! + \log (T-t_i)! - \log T! \\ & - \log k! - \log (f-k)! - \log (t_i-k)! - \log (T-f-t_i+k)!\end{aligned}$$

Figure 3. Formula for calculating probability in a text corpus

Four parameters can vary: the frequency f of a word throughout the corpus, its frequency k in the section, the corpus size T and the section size t_i (Labbé & Labbé 2001). Logarithms can be used for easier computing in corpus analysis, because corpus size and frequency result in high numbers, mounting up to thousands for frequency and even several millions for corpus size. Using logarithms in the right side of the formula with factorials implies that the result of the formula is the log of the probability and not the probability itself. The result log Prob($X=k$) = y is to be interpreted as the exponent base 10, leading to a probability of 10^y.

When this log formula is applied to our example of red balls, Prob($X=6$) equals log14!+log12!–log20!–log2!–log4! or −0.4466. As mentioned above, formulated as an exponent base 10, this results in the probability $10^{-0.4466}$ or 0.3576.

Note that in a very large text corpus, hypergeometric distribution is sometimes approximated by binomial, Poisson or normal distribution, in order to calculate the probability of high frequencies.

2.3 Results of Calculating Specificities: S^+ and S^-

The Calculation of Specificities yields a value which indicates whether the frequency of a word in a section is typical or not. For the text section *i*, the probability Prob ($X = k$) reaches a maximum at the expected frequency in the part *i* (f'_i): the word occurs as many times as expected, according to its frequency throughout the text corpus. Whenever the observed frequency does not equal the expected frequency, it is important to check in how far the difference between the observed and expected frequency is significant (Labbé & Labbé 2001). If the observed frequency is higher than the expected frequency, the calculated probability of the observed frequency Prob ($X = k$) will be S^+ = Prob ($X \geq k$). If the probability is smaller than a threshold value defined at 0.05 or 0.01, the word is assigned 'positive specificity'[2] (Lafon 1984). The word appears more often in the section than expected by chance or more often than expected when comparing to its frequency in the entire corpus. Conversely, when the observed frequency is lower than the expected frequency, the calculated probability of the observed frequency Prob($X = k$) will be $S^- = \text{Prob}\,(X \leq k)$. If this probability is smaller than a defined threshold value, the word is assigned 'negative specificity' (Lafon 1984), meaning it appears less often than expected by chance. Since positive keywords appear significantly more often in the analysed section, some of them can be recognised through reading the section. However, negative keywords only appear when contrasted with the entire corpus.

2.4 Probability value and typicality coefficient value

The outcome of the formula for hypergeometric distribution indicates the probability of the observed frequency as compared to the expected frequency. Since high word frequencies and corpus size of textual data require logarithms of the factorials, the outcome is expressed through the logarithm of the probability, which is the exponent base 10 to achieve the probability value as a result. This exponent, called the 'typicality coefficient value', is the only outcome of the formula for hypergeometric distribution. Mathematically, the formula yields one result. There is no associated test statistic, in contrast to the results of the Keywords Method (see Section 3.2). A high value for this exponent means a low probability for the

observed frequency. For example, for the result log Prob($X = k$) = −2, Prob($X = k$) equals 10^{-2} or 0.01 (1%). The higher the exponent, the lower the probability of the word's frequency in the section and the more key the word. Therefore, there is a positive correlation between keyness and the exponent, being indirectly the probability value. As explained before, the exponent can be considered as the word's typicality coefficient value and, furthermore, put as an exponent to base 10, it shows the probability of obtaining the word's frequency.

3. *WordSmith Tools* and the underlying Keywords Method

Mike Scott developed *WordSmith Tools* (or WS) for corpus analysis at the University of Liverpool (1996). The most recent version is *WordSmith Tools* 6.0 (Scott 2013). The software includes the applications WS WordList for frequency lists of words and word clusters, WS Concord for concordances and collocation analysis, WS Concgram to derive concgrams, as well as WS KeyWords for keyword identification.[3]

3.1 *WordSmith Tools*

WS KeyWords, based on the underlying methodology of Keywords Method, enables extraction of keywords in a text or corpus. Such keywords are often interpreted as indication of 'aboutness' (Baker 2004; Scott 2013). WS also enables the identification of 'key keywords' (keywords which are key in a number of text files) (Scott 2013), 'associates' (keywords which are commonly associated with a key keyword, because they are key in the same texts as the key keyword) and 'linked keywords' (keywords which co-occur within a collocational span) and thus keyword clusters; for example, the keyword *Jesus* and the keyword *Christ* are very often found near each other and thus they form the keyword cluster *Jesus Christ*. The extracted keywords can also be plotted in order to visualise how often and where they crop up in the text, and how they are dispersed or clustered.

Keywords occur significantly more often in a text or corpus, for example a specialist or technical corpus, than one would predict or expect on the basis of their frequency in the language as a whole, represented by a reference corpus, for example a general language corpus (Gries 2006; Scott 2009). WS KeyWords cross-tabulates the frequency of a word in a word list derived from the specialist corpus, the number of running words in this word list, its frequency in the reference word list and the number of running words in the reference word list. Statistical tests include the classic χ^2 test of significance with Yates correction and the log-likelihood test statistic (LLR) (Cressie & Read 1989; Dunning 1993) (for details

on LLR, see Appendix 2). Several settings can be modified including probability value, minimal word frequency and the statistical test (χ^2 or LLR). A word is considered a keyword if its frequency is either unusually high (positive keyword) or unusually low (negative keyword), in comparison with what would be expected on the basis of the reference word list.

In order to determine the keywords of a technical corpus, two word frequency lists are loaded in WS WordList, one for the technical corpus and another for the reference corpus. In the tool WS KeyWords, a frequency list and a reference frequency list are selected in order to create a keyword list. The default *p*-value is 0.000001, which means that only the most typical keywords are identified. The results can be saved and imported in Excel. They are represented in 8 columns (see Table 2) which include rank order of the keyword, keyword, absolute and relative frequency in the technical corpus, absolute and relative frequency in the reference corpus, keyness (i.e. the statistical measure) and *p*-value.

Table 2. Output WS KeyWords: keywords information in 8 columns

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	*p*
1	usinage	559	0,6111	0		2552,1	5,01038E-20
2	machine	529	0,5783	83	0,0103	1946,9	1,14283E-19
3	outil	476	0,5204	67		1781,4	1,49936E-19
4	axe	398	0,4351	12		1710,6	1,69743E-19
5	mm	395	0,4318	13		1690,3	1,76054E-19
6	pièce	518	0,5663	154	0,0192	1674,3	1,81256E-19
7	x	351	0,3837	22		1439,3	2,88275E-19
8	Fig	259	0,2831	0		1181,7	5,29191E-19
9	broche	256	0,2799	0		1168	5,48597E-19
10	vitesse	266	0,2908	31		1021,6	8,30587E-19
11	un	4101	4,4833	21360	2,6606	859,31	1,4231E-18
12	précision	203	0,2219	27		765,52	2,04371E-18
13	diamètre	154	0,1684	2		681,5	2,94695E-18
14	plaquette	146	0,1596	1		654,21	3,35335E-18
15	Trametal	140	0,1531	0		638,59	3,62E-18
16	jusque	168	0,1837	27		615,33	4,07219E-18
17	type	188	0,2055	82	0,0102	543,74	6,04004E-18
18	tournage	135	0,1476	11		540,11	6,17095E-18
19	acier	126	0,1377	5		533,33	6,42594E-18
20	m/mn	113	0,1235	0		515,4	7,17074E-18
21	système	248	0,2711	254	0,0316	490,31	8,41982E-18
22	t/mn	105	0,1148	0		478,91	9,08418E-18
23	machine-outil	103	0,1126	0		469,78	9,6668E-18
24	coupe	172	0,188	93	0,0116	461,19	1,02621E-17

3.2 The underlying Keywords Method

In contrast to the Calculation of Specificities, the Keywords Method or the keywords procedure (Scott 2001) does not involve a part-whole comparison. A technical corpus (LSP) is compared to a reference corpus (LGP) in order to identify words that are typical of the specialised domain. Since the Keywords Method compares frequencies in two different corpora, the data can be presented in a contingency table (see Table 3). Relative frequency of a word expresses the ratio between absolute frequency of that word (i.e. observed frequency) and corpus size (i.e. the total number of words in that corpus), for example a/N_1 or k/t for the LSP corpus and b/N_2 for the LGP reference corpus (see Figure 1 for codes *k, t, f* and *T* from Section 2.2.1). In order to easily compare the Keywords Method and the Calculation of Specificities, we also consider a hybrid corpus, consisting of both the LSP and the LGP corpus. For the frequency *a* in the LSP corpus and the frequency *b* in the LGP corpus, we consider the frequency *f* throughout the hybrid LSP+LGP corpus.

Table 3. Contingency table for frequencies of a word in two corpora

	LSP corpus	**LGP corpus**	**Total = LSP+LGP corpus**
Frequency of a word	$a\ (= k)$	b	$a + b = f$
Corpus size	$N_1\ (= t)$	N_2	$N_1+N_2 = T$

In contrast to absolute frequencies, relative frequencies allow easy data comparison between two corpora. If the observed relative frequency of a word in the target (LSP) corpus equals its expected relative frequency, the word occurs as frequently in the target corpus as in the reference corpus, which means that it is not a keyword. In cases where its observed relative frequency in the target corpus exceeds its expected relative frequency, the word is considered to be a keyword, typical for the target corpus.

In order to calculate the LLR value of a word, we need its frequency in both the technical corpus (LSP) and the reference corpus (LGP) as well as the size of both corpora, as shown in Table 4. Note that $\neg a$ stands for "not *a*".

Table 4. Contingency table for word frequency comparison

	LSP corpus	**LGP corpus**	**Total = virtual corpus**
Frequency of a word	a	b	$a + b$
Frequency of the other words	$\neg a\ (N_1-a)$	$\neg b\ (N_2-b)$	$\neg a + \neg b$
Corpus size	N_1	N_2	$N\ (N_1+N_2)$

The LLR value is calculated as shown in Figure 4, with $E_1 = N_1*(a+b)/(N_1+N_2)$ for the specialised LSP corpus and E_2 by analogy for the reference corpus or LGP corpus (Rayson & Garside 2000: 3). In Figure 4, the log is to base *e* or ln (natural log).

$$LLR = 2*((a*log(a/E_1)) + (b*log(b/E_2)))$$

Figure 4. Calculation of the LLR value

3.3 Results of the Keywords Method

The Keywords Method yields a LLR value, which indicates the degree of typicality of a keyword, after statistical comparison of the frequencies in the technical and the reference corpus. The corresponding *p*-value allows a cut-off point for significance, for example $p < 0.05$ (for LLR ≥ 3.84).

4. Comparative study

A comparative experiment was conducted on a small test corpus. The LSP part consists of technical French texts from the machining domain (100,168 tokens). The reference part comprises newspaper articles from *Le Monde* and counts about 900,000 tokens. Both parts are lemmatised and tagged with the tool *Cordial 7 Analyseur* (Audibert 2002).[4] We use lemmas instead of individual word forms in order to consider all occurrences of the same lemma as one keyword (e.g. singular and plural noun forms, male and female adjectives in French, inflected verb forms). The two parts constitute the complete LSP+LGP corpus of about one million tokens. If the technical part is compared to the complete corpus in the tool L3, the part-whole ratio is about 1/10 (see Section 2.1). To facilitate the comparison of the results, *p*-value is set at 0.05 in WS. The minimal frequency is 1, i.e. all words are included in the keyword analysis. The option 'hyphens break words' in the text characteristics settings must be disabled in order to maintain words with hyphen (e.g. *machine-outil*) and 'characters within words' are set as follows '–/'.

On the whole, the tools used in our comparative study generate relatively similar results for extracted keywords and their ranking (see Section 4.1). Nevertheless, the experiments also show important differences with regard to probability value, typicality coefficient and reference corpus (see Section 4.2). As mentioned in the methodological sections, some of these differences are direct consequences of the underlying methodology and statistics.

4.1 Similarities between the keyword lists

WS indicates positive keywords in colour or allows in its settings to exclude negative keywords. In L3 positive keywords are shown separately by means of an extra option in the results window (see Table 1). A global comparison of the output shows differences with respect to the number of positive keywords detected in the LSP corpus (see Table 5). There are also smaller differences with respect to the total number of types, which can be explained by the way the two tools consider words, signs, figures, etc. to be types.

Table 5. Number of positive keywords and number of types in LSP corpus

	L3	WS	WS including figures
Number of positive keywords	4,459	2,665	3,113
Number of types in LSP corpus	5,935	5,720	5,860

The L3 list is much longer (4,459 keywords) than the WS list without figures (2,665 keywords). This difference is partly due to the figures included in the L3 list, but the second WS list (including figures), with 3,113 keywords, reveals that the L3 list is still longer. The difference in length may be explained in terms of probability value because the end of the L3 list has a coefficient of 1. As will be explained in Section 4.2.1, this probability value is not always statistically significant. As a consequence, the words at the end of the L3 list are not all statistically significant keywords.

A comparison of the keywords actually extracted produces similar results, as shown in Table 6, for the 30 most typical keywords of both keyword lists. They belong to the same word classes and are mostly nouns. As for the ranking of keywords, WS uses the fine-grained LLR statistic, called 'keyness' in the output table. Ranking differences observed in the L3 list are mainly due to the fact that the same typicality coefficient is attributed to several keywords (see Section 4.2.2). Moreover, the 62 most typical keywords in L3 all have three asterisks (***) in the column of the typicality coefficient, instead of an exact value. Unfortunately, this does not allow for a comparison in terms of rank correlation statistics.

If the two keyword lists are sorted by descending frequency in the specialised part (L3) or the specialised corpus (WS), the results are very similar, as shown in Table 7, despite some small differences. 24 keywords out of 30 appear in both WS and L3 lists at similar ranks (Spearman's rank correlation: 0.76).

A comparison of the full lists in both tools, i.e. 4,459 keywords in L3 and 2,665 in WS, shows more similarities with respect to the extracted keywords than differences. Almost all WS keywords are found in the L3 list, except for the

Table 6. Top 30 positive keywords in WS and L3

	WS	Keyness	L3	Coefficient
1	*usinage*	2552.101	*usinage*	***
2	*machine*	1946.948	*machine*	***
3	*outil*	1781.411	*pièce*	***
4	*axe*	1710.612	*outil*	***
5	*mm*	1690.333	*axe*	***
6	*pièce*	1674.334	*mm*	***
7	*x*	1439.274	*x*	***
8	*Fig*	1181.693	*permettre*	***
9	*broche*	1167.998	*vitesse*	***
10	*vitesse*	1021.571	*Fig*	***
11	*un*	859.3113	*broche*	***
12	*précision*	765.515	*système*	***
13	*diamètre*	681.4992	*centre*	***
14	*plaquette*	654.2074	*précision*	***
15	*Trametal*	638.5892	*type*	***
16	*jusque*	615.3273	*modèle*	***
17	*type*	543.7416	*2000*	***
18	*tournage*	540.1111	*assurer*	***
19	*acier*	533.3264	*coupe*	***
20	*m/mn*	515.4028	*jusque*	***
21	*système*	490.3125	*diamètre*	***
22	*t/mn*	478.9058	*plaquette*	***
23	*machine-outil*	469.7818	*Trametal*	***
24	*coupe*	461.1867	*tournage*	***
25	*permettre*	459.0388	*acier*	***
26	*usiner*	446.972	*0*	***
27	*fraise*	440.5492	*moteur*	***
28	*modèle*	410.6217	*rapide*	***
29	*linéaire*	409.7129	*gamme*	***
30	*moteur*	396.6062	*m/mn*	***

Table 7. Keywords in WS list not appearing in L3 list

N	Keyword	LSP fq	LGP fq	Keyness	*p*-value
51	#	3413	21924	275.0108	5.71E-17
1206	LX	2	0	9.1199	2.53E-03
1843	RV	1	0	4.5599	3.27E-02
2056	FH	1	0	4.5599	3.27E-02
2374	PRES	1	0	4.5599	3.27E-02
2460	IA	1	0	4.5599	3.27E-02

6 keywords shown in Table 7. These are not in the L3 list, due to zero frequency in the LGP corpus. If using a frequency cut off, e.g. minimal frequency 3, they would not appear in the WS list.

We have already discussed the keyword #. The other keywords are very rare items; they appear only once or twice in the specialised corpus. They are not typical keywords, because they have a LLR value (keyness) of 9 or 4, which is of borderline significance.

The comparison of the L3 list to the WS list yields a longer list of unmatched items: 1,795 keywords in L3 are not detected in WS. As mentioned before, most of these unmatched keywords include numbers or have a typicality coefficient of 1 ($p < 0.1$), which means that they are not always statistically significant. The other unmatched keywords are more general words, e.g. *novembre*, *mai*, *index*, *croix*, *mettre*, *dernier*, *donner*, *différer*. It is clear that L3 selects more general words than WS, even among the statistical significant keywords with typicality coefficient > 1.

Once the keywords are extracted, function words (e.g. *de*, *un*, *à*, *et*, *en*, *pour*) and proper nouns (e.g. *Trametal*) can be filtered out. Typicality coefficients are then used to sort keywords by descending degree of typicality and to position them on the typicality continuum, ranging from most typical to least typical keywords. The most typical keywords (*machine*, *outil*, *usinage*, *pièce*, *mm*, *vitesse*, *coupe*) clearly reflect the thematic content of the machining domain. Note that the extracted keywords are not only terms in the strict sense, such as *usinage* and *fraisage*, but also general language words (e.g. *type*, *permettre*) very often used in technical writing.

4.2 Differences between the keyword lists

Differences between the outcome of the tools fall within the parameters of probability value, typicality coefficient and reference corpus. The probability value and the typicality coefficient are always associated and depend on the underlying statistical measure. The differences with respect to the reference corpus are mainly due to the methodological choice (part-whole versus specialised-general).

4.2.1 *Probability value*

p-value in WS is set at 0.05, which is statistically significant, but not very severe. For keyword identification, L3 shows in the application Spécificités a window with 2 parameters, namely *p*-value and minimal frequency. *p*-value possibilities range from 1 to 100. In L3, the probability value is also obtained indirectly via the typicality coefficient, which is an integer number (see Section 4.2.2).

In order to refine the comparison of L3 and WS, we compare the number of positive keywords at more severe significance levels of 0.01, 0.001 and even 0.0001. Table 8 shows that the results of L3 and WS are more convergent for the more significant and thus more typical keywords ($p < 0.001$ and 0.0001). Remaining differences are mainly due to the underlying methodology of part-whole comparison in L3 (see Section 4.2.3).

Table 8. Number of positive keywords in the 2 tools for 3 significance levels

	$p < 0.01$	$p < 0.001$	$p < 0.0001$
L3	1,928	1,275	954
WS	1,539	1,093	829
WS with figures	1,707	1,187	899

4.2.2 *Typicality coefficient*

In WS, the probability value of a keyword is associated with a typicality coefficient, indicating the keyness of that keyword. Keywords can be sorted by descending keyness in order to rank them. This kind of sorting generates a typicality ranking from most typical to less typical, but all keywords in the WS list are statistically significant. However, in a list including several thousands of keywords, some of them not only have the same relative frequency in the target corpus but also the same relative frequency in the reference corpus. As a consequence, these words will have the same LLR value (typicality coefficient) and thus the same degree of typicality. This does not have any influence on the linguistic interpretation of the typicality, since these words share the same frequency characteristics.

In L3, the results do not include a column for a test statistic, because there is no test statistic. From a mathematical point of view, calculating the result of the underlying exact hypergeometric distribution means calculating the log of the probability. This result is implemented in L3 in terms of a typicality coefficient ("coefficient de spécificité"). Keywords can be ranked when sorted by the typicality coefficient, which is an integer between 1 and 50; three asterisks (***) are used to indicate the most typical keywords. It is clear that the integer does not allow as fine-grained distinctions as the LLR value in WS. The larger the corpora compared, the more keywords receive the same coefficient, which impedes a clear-cut ranking with regard to the degree of typicality. In a previous experiment with a very small specialised corpus of only 14,000 tokens, the 10 most typical words all have a coefficient value of 50. The specialised corpus of this experiment is made up of 100,000 tokens and reveals 62 keywords with *** indication, 2 keywords with typicality coefficient value 50 and 2 keywords with value 49 (see Table 9). For the lowest typicality coefficient values (2 and 3), the granularity problem is

even more complex: in the list of 4,459 keywords, 2,052 keywords have typicality coefficient 2 and 653 keywords have typicality coefficient 3.

Table 9. Top 70 positive keywords in L3

Keyword	Total fq	Fq part	Coeff.	Keyword	Total fq	Fq part	Coeff.
diamètre	156	154	***	–	901	648	***
course	156	112	***	*modèle*	342	186	***
moteur	150	119	***	*m/mn*	113	113	***
gamme	154	115	***	*géométrie*	72	68	***
constructeur	161	107	***	*mandrin*	69	69	***
0	221	120	***	*taraudage*	69	69	***
précision	230	203	***	*automatique*	76	67	***
rapide	178	116	***	*copeau*	81	81	***
jusque	195	168	***	*fraisage*	77	77	***
déplacement	125	90	***	*kw*	77	77	***
conception	131	105	***	*à*	18064	2657	***
surface	119	84	***	*un*	25461	4101	***
commande	124	84	***	*de*	75360	9176	***
acier	131	126	***	*carbure*	55	55	***
tournage	146	135	***	*vertical*	68	65	***
plaquette	147	146	***	*rotatif*	64	64	***
fabrication	133	104	***	*finition*	61	57	***
Trametal	140	140	***	*usiner*	98	98	***
broche	256	256	***	*t/mn*	105	105	***
système	502	248	***	*utilisateur*	97	73	***
centre	513	225	***	*fraise*	100	99	***
mm	408	395	***	*machine-outil*	103	103	***
axe	410	398	***	*nuance*	106	84	***
outil	543	476	***	*avance*	111	84	***
machine	612	529	***	*linéaire*	96	94	***
pièce	672	518	***	*z*	95	85	***
usinage	559	559	***	*taraud*	90	89	***
permettre	804	313	***	*fonte*	57	54	50
type	270	188	***	*arête*	52	51	50
2000	281	177	***	*application*	155	91	49
Fig	259	259	***	*équiper*	135	84	49
coupe	265	172	***	*revêtement*	53	51	48
vitesse	297	266	***	*très*	1000	266	48
x	373	351	***	*usure*	51	49	46
assurer	399	173	***	*contrôle*	210	103	46

4.2.3 *Reference corpus*

As described in the methodological Sections 2.1 and 3.1, there are two types of reference corpus used in this study: (i) part-whole comparison and (ii) comparison of the specialised corpus to a general language reference corpus. The first type of reference corpus is used with the Calculation of Specificities method discussed in Section 2.2. Applying the part-whole comparison to our data requires incorporation of our technical corpus (LSP) and our general language corpus (LGP) into one large virtual reference corpus (LSP+LGP corpus). From a methodological point of view, this incorporation is not very satisfying, since the reference corpus becomes heterogeneous, having one specific section and nine general sections. The second type of reference corpus is used with the Keywords Method discussed in Section 3.2 which compares two independent corpora. The reference corpus is a general language corpus and does not include the LSP corpus.

L3 does not allow the comparison of one corpus to another corpus, only the comparison of one part to the entire corpus (part-whole comparison). By way of experiment, we also implement the part-whole comparison in WS (see Table 10), in order to see what that means for the number of extracted keywords. We note that the implementation of the part-whole comparison in WS is not methodologically correct, because the underlying LLR test statistic is not suited for a part-whole comparison.

Table 10. Positive keywords in L3 and WS (various configurations of reference corpus)

		$p < 0.05$ all frequencies
L3	(LSP corpus vs. LSP+LGP corpus)	4,459
WS	(LSP corpus vs. LGP corpus)	2,665
WS	(LSP corpus vs. LSP+LGP corpus)	1,600

This keyword list contains fewer positive keywords in WS (1,600), whereas the regular keyword list counts 2,665 keywords, suggesting that the part-whole comparison (LSP versus LSP+LGP) dilutes the number of positive keywords obtained. This dilution can be explained by the fact that in the combined LSP+LGP corpus, the total frequency of the words is higher than their frequency in the methodologically sound reference corpus (LGP). Since all words appear at least once in this entire heterogeneous corpus (LSP+LGP), fewer words will be "key" in the LSP corpus.

5. Conclusions

Both tools, L3 and WS, can be used for keyword extraction and yield a similar list of keywords with a typicality coefficient. In view of our specific research goals and research corpora, the Keywords Method seems technically more efficient for establishing a fine-grained typicality ranking of all the keywords of a technical corpus with respect to a general language reference corpus. Determining factors governing the choice of tool include the use of a general language reference corpus and the granularity of the typicality coefficient value. L3 is mainly used to identify the typical vocabulary or keywords of one section in comparison with the entire corpus, whereas WS is generally used for keyword extraction in a target corpus with respect to a reference corpus. Since our study deals with keyword identification in a technical corpus, as opposed to a general language reference corpus, the Keywords Method and WS are most appropriate.

The global approach adopted in both tools is similar. Nonetheless, the scale of deviation is different, opposing a hypergeometric distribution (Calculation of Specificities) to an asymptotic distribution for the LLR statistic (Keywords Method). The hypergeometric distribution seems to be less appropriate for the analysis of extensive corpora in that it cannot statistically handle high frequencies. This is due to the calculation rather than the distribution itself. L3 is most suited for relatively small text documents, where one section is compared to the entire document. The column with the typicality coefficient allows cutting off according to various *p*-value thresholds. Unfortunately, L3 very often attributes the same coefficient value to several keywords at a time and, therefore, a fine-grained ranking of keywords is impossible. Our research requires a well-established typicality continuum with a fine-grained ranking range. However, if a fine-grained typicality ranking is not required, L3 provides clear and reliable results. It is important to note that the package L3 consists of several very useful tools for lexicometric analysis and textual statistics and that the keywords extraction tool is not one of the main tools. In WS, large corpora do not pose a processing problem. The probability value can be modified and the corresponding column in the results allows cutting off at lower *p*-value thresholds. Furthermore, the keyness value allows a very fine-grained typicality ranking.

In this paper, we focused on single word items, since it is difficult to determine the typicality degree of multiword expressions in a technical domain using the Keywords Method. While extraction of multiword items has already been carried out in *Wmatrix* (Rayson 2009) for general language, most technical multiword expressions do not appear in a general language reference corpus. However, our study has implications for further research which moves beyond the lexical level. Several studies already deal with key clusters (see Baker 2004). Key clusters

are recurrent sequences whose frequencies can be compared using for example the LLR test statistic. The problem here is one of using the same measure for words as for clusters.

Notes

1. The comparative study was carried out in the context of PhD research that set out to analyse the typical vocabulary (or keywords) of a specialised technical corpus in French. Building on a quantitative approach and corpus data, this investigation attempted to find out to what extent keywords in a technical domain are monosemous or polysemous. As a consequence, all keywords of the technical corpus needed to be extracted, with their degree of keyness or typicality, as fine-grained as possible.

2. Note that the sign + or – of S is not part of the hypergeometric calculation. It is added after the calculation in order to distinguish between 'positive' and 'negative' keywords.

3. For a list of examples of research using *WordSmith Tools* see http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm (accessed September 2013).

4. Synapse Développement Editeur de logiciels: http://www.synapse-fr.com/ (accessed September 2013).

References

Audibert, L. 2002. "Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences". In *Actes de RECITAL (TALN) 2002*, 415–424.

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

Baron, A., Rayson, P. & Archer, D. 2009. "Word frequency and key word statistics in corpus linguistics". *Anglistik: International Journal of English Studies*, 20 (1), 41–67.

Berber Sardinha, A. 1996. "Review: WordSmith Tools". *Computers & Texts*, 12, 19–21.

Berber Sardinha, A. 1999a. "Word sets, keywords and text contents: An investigation of text topic on the computer". *DELTA*, 15 (1), 141–149.

Berber Sardinha, A. 1999b. "Using KeyWords in text analysis: Practical aspects". *DIRECT Papers*, 42, 1–8.

Bertels, A. 2005. "A la découverte de la polysémie des spécificités du français technique". In *Actes de TALN* et *RECITAL 2005*, 575–584.

Bertels, A., Speelman, D. & Geeraerts, D. 2006. "Analyse quantitative et statistique de la sémantique dans un corpus technique". In *Actes de TALN 2006,* 73–82.

Bondi, M. & Scott, M. 2010. *Keyness in Text*. Amsterdam: John Benjamins.

Cressie, N., & Read, T. R. C. 1989. "Pearson's X2 and the log-likelihood ratio statistic G2: A comparative review". *International Statistical Review*, 57 (1), 19–43.

Drouin, P. 2003. "Term extraction using non-technical corpora as a point of leverage". *Terminology*, 9 (1), 99–117.

Drouin, P. 2004. "Spécificités lexicales et acquisition de la terminologie". In *Actes de JADT 2004*, 345–352.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics*, 19 (1), 61–74.

Evert, S. & Krenn, B. 2001. "Methods for the qualitative evaluation of lexical association measures". *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France*, 188–195.

Evert, S. & Krenn, B. 2003. "Computational approaches to collocations". *Introductory Course at ESSLLI 2003*. Available at: http://www.collocations.de/EK/index.html (accessed September 2013).

Evert, S. 2002. "Special topic session on the mathematical properties of association measures". *Presentation at the Workshop on Computational Approaches to Collocations. Vienna*. Available at: http://www.collocations.de/EK/index.html (accessed July 2013).

Evert, S. 2007. *Corpora and Collocations*. Extended Manuscript of Chapter 58 of Lüdeling A. & M. Kytö, 2008, *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin. Available at: http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf (accessed October 2013).

Fonseca-Greber, B. B. 2007. "The emergence of emphatic 'ne' in conversational Swiss French". *Journal of French Language Studies*, 17 (3), 249–275.

Gries, S. T. 2006. "Exploring variability within and between corpora: Some methodological considerations". *Corpora*, 1 (2), 109–151.

Labbé, C. & Labbé, D. 2001. "Que mesure la spécificité du vocabulaire?" *Lexicometrica* 3. Available at:http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3/specificite2001.PDF (accessed September 2013).

Lafon, P. 1984. *Dépouillements et Statistiques en Lexicométrie*. Genève–Paris: Slatkine-Champion.

Lamalle, C. & Salem, A. 2002. "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels". In *Actes de JADT 2002*, 403–412.

Lebart, L. & Salem, A. 1994. *Statistique Textuelle*. Paris: Dunod.

Lemay, C., L'Homme, M. C. & Drouin, P. 2005. "Two methods for extracting specific single-word terms from specialized corpora. Experimentation and evaluation". *International Journal of Corpus Linguistics*, 10 (2), 227–255.

Lyne, A. A. 1985. *The Vocabulary of French Business Correspondence*. Geneva: Slatkine.

Manning, C. & Schütze, H. 2002. *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press.

Martinez, W. 2000. "Mise en évidence de rapports synonymiques par la méthode des cooccurrences". In *Actes de JADT 2000*, 78–84.

Moore, R. C. 2004. "On log-likelihood-ratios and the significance of rare events". Available at: http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Moore.pdf (accessed September 2013).

Müller, C. 1992 [1977]. *Principes et Méthodes de Statistique Lexicale*. Paris: Champion.

Poibeau, T. 2004. "Pré-analyse de corpus". In *Actes de JADT 2004,* 897–903.

Rayson, P. & Garside, R. 2000. "Comparing corpora using frequency profiling". In *Proceedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, 1–6.

Rayson, P. 2008. "From key words to key semantic domains". *International Journal of Corpus Linguistics*, 13 (4), 519–549.

Ross, S. 1994. *A First Course in Probability*. New York: Macmillan College Publishing Company.
Salem, A. 1987. *Pratique des Segments Rrépétés: Essai de Statistique Textuelle*. Paris: Klincksieck.
Sansonetti, L. 2003. "Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultats d'expérience". In *Actes des 3èmes Journées de Linguistique de Corpus 2003* (71–85). Available at: http://web.univ-ubs.fr/corpus/jlc3/1_4_sansonetti.pdf (accessed September 2013).
Scott, M. & Tribble, C. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Scott, M. 2000. "Focusing on the text and its key words". In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective, Volume 2*. Frankfurt: Peter Lang, 103–122.
Scott, M. 2001. "Mapping key words to problem and solution". In M. Scott & G. Thompson (Eds.), *Patterns of Text: In Honour of Michael Hoey*. Amsterdam: Benjamins, 109–127.
Scott, M. 2009. "In Search of a bad reference corpus". In D. Archer (Ed.), *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate, 79–92.
Stubbs, M. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies". *Functions of Language*, 2 (1), 23–55.
Tribble, C. 1999. "Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals". In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective, Volume 2*. Frankfurt: Peter Lang, 75–90.
Vangehuchten, L. 2004. "El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos: Descripción del proceso de selección del léxico típico del discurso económico empresarial en español". In *Actes de JADT 2004*, 1128–1135.
Weber, M., Vos, R. & Baayen, H. 2000. "Extracting the lowest-frequency words: Pitfalls and possibilities". *Computational Linguistics*, 26 (3), 301–317.
Williams, G. 2002. "In search of representativity in specialised corpora. Categorisation through collocation". *International Journal of Corpus Linguistics*, 7 (1), 43–64.
Zimina, M. 2004. "Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles". In *Actes de JADT 2004*, 1195–1202.

Tools

AntConc

Anthony, L. 2011. *AntConc* (Version 3.2.2). Tokyo, Japan: Waseda University. Available at: http://www.antlab.sci.waseda.ac.jp/ (accessed September 2013).

AV Frequency List Tool

Speelman, D. 1997. *Abundantia Verborum: A Computer Tool for Carrying out Corpus-based Linguistic Case Studies*. Unpublished PhD Thesis. Leuven, Belgium: KULeuven. Available at: http://wwwling.arts.kuleuven.be/qlvl/ToolsTraining.htm (accessed September 2013).

Lexico 3

Lamalle, C., Martinez, W., Fleury, S. & Salem, A. 2003. *Outils de Statistique Textuelle. Manuel d'Utilisation de Lexico3*. Paris: Université de Paris3. Available at: http://www.tal.univ-paris3.fr/lexico/ (accessed September 2013).

Wmatrix

Rayson, P. 2009. *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster University. Available at: http://ucrel.lancs.ac.uk/wmatrix/ (accessed September 2013).

WordSmith Tools

Scott, M. 2013. *WordSmith Tools. Version 6*. Liverpool: Lexical Analysis Software. Available at: http://www.lexically.net/wordsmith/ (accessed September 2013).

Appendix

1. Hypergeometric distribution

A hypergeometric distribution is a discrete probability distribution and describes the number of successes in a sequence of n draws (fixed size sample) from a discrete and finite population, without replacement. For example, in a population of N balls, m balls are red and the others (N–m) are white. If we select and withdraw n balls from the population, what is the probability that exactly k selected balls are red? The distribution of the red balls among the n selected balls follows a hypergeometric distribution (see Figure 5). The formula for hypergeometric distribution in Figure 5 is a fraction. The denominator is a binomial coefficient indicating the number of possible arrangements of the n selected balls in the population of N balls. The numerator gives all the possible ways of arranging the red and white balls respectively.

$$\text{Prob}\,(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Figure 5. General formula for hypergeometric distribution

To be more precise, in a box with 20 balls, 14 balls are red and 6 balls are white. If 8 balls are randomly extracted, what is the likelihood that exactly 6 selected balls are red? When applying the general formula for hypergeometric distribution (see Figure 5), the probability that exactly 6 red balls occur within 8 balls taken out of 20 is as follows (see Figure 6). In Figure 6 the numerator and the denominator of the general formula are filled in with the concrete data of these red and white balls.

$$\text{Prob}\,(X=6) = \frac{\binom{14}{6}\binom{20-14}{8-6}}{\binom{20}{8}}$$

Figure 6. Sample formula for hypergeometric distribution

Hypergeometric distribution is characterised by the lack of replacement. For a sample of 8 balls out of 20, the probability of these being red changes with every ball extracted, because not only the total number of balls changes, but also the distribution of red and white balls. These changes in probability are accounted for in the formula by including both the possibilities for the red part and the white part as well as the possibilities for the whole sample. For a sample of 8 balls out of 20 balls, there are $\binom{20}{8}$ possibilities, which means $\frac{20!}{(20-8)!8!} = 125.970$ possibilities. $\binom{20}{8}$ is called a binomial coefficient and expresses all the possible ways of arranging each combination. As a consequence, each chance or selection of 8 balls has the value of $1/\binom{20}{8}$. For a sample of 6 red and 2 white balls out of a population of 14 red and 6 white balls, there are $\binom{14}{6}$ and $\binom{20-14}{8-6}$ or $\binom{6}{2}$ possibilities. There are many feasible samples of 8 balls out of 20 without replacement, equally, there are many ways to obtain 6 red balls just as there are many ways to fill in the rest of the sample with 2 white balls. The product of these is given by the formula in Figure 8. The result of the product is the probability that exactly 6 balls in a sample of 8 balls out of 20 will be red. In this example, the result of Prob($X = 6$) equals 0.3576 (or 35.76%).

2. Log-likelihood ratio

When comparing data in a contingency table, several statistics can be used, for example Pearsons's chi-squared (χ^2) test statistic (Manning & Schütze 2002), Z-score or Mutual Information (MI) (Church & Hanks 1990). MI and χ^2 seem to overestimate the significance of low frequency events. This overestimation problem is well-known and we refer the interested reader to Church & Gale (1991b), Stubbs (1995), Weber et al. (2000), Evert & Krenn (2001) and Manning & Schütze (2002). The underlying assumption of Z-score is a normal distribution (Evert 2002), which supposes that the events being analysed are relatively common. In order to overcome such overestimation problems, the log-likelihood test statistic (LLR or G^2), first introduced into statistics by Dunning (1993), is a good alternative, especially when counts are small (smaller than 20 or between 20 and 40) and if the expected value is 5 or less (Manning & Schütze 2002). This general test statistic is efficient for both small and large corpora and allows a direct comparison of the significance of rare and common events. It does not assume a normality distribution for the word frequencies, but an asymptotic or approximative χ^2 distribution. As a consequence, the significance of rare words is more reliable.

The LLR test statistic is used to check the independence of two variables, for example *a* and *b* in Table 3, i.e. the observed frequency in a specialised corpus LSP and the observed frequency in a reference corpus LGP. In fact, the data values of both corpora, as illustrated in the first two columns of the contingency table (see Table 3), are considered as two different samples and are checked as to whether those two samples are drawn from the same population. The null hypothesis postulates that *a* and *b* come from the same population and thus have the same frequency distribution. An alternative hypothesis states that *a* and *b* come from different populations and differ significantly. This alternative hypothesis is accepted if the observed data are sufficiently improbable given the null hypothesis. The probability can be calculated by means of a significance test, such as LLR. The LLR test statistic is based on the ratio L of two likelihoods: (i) in the numerator, the maximum likelihood given the null hypothesis, in which two samples are drawn from the same population and (ii) in the denominator, the overall maximum

likelihood. According to the null hypothesis, a word has the same frequency distribution in the specialised and in the reference corpus. This means that within the null hypothesis, its observed frequency in the specialised corpus almost equals its expected frequency (determined on the basis of its observed frequency in the reference corpus). However, for keywords of the specialised corpus, there is a significant difference between their frequency in the specialised and in the reference corpus, meaning that the null hypothesis is false. As a consequence, for keywords, the probability of the outcome will be very low and the null hypothesis (no frequency or distribution difference) can be rejected.

The ratio L is a number between 0 and 1. The closer this number approaches 0, the lower the maximum likelihood of the numerator given the null hypothesis, which is a strong indication that the null hypothesis is false. The actual LLR test statistic or log-likelihood ratio, however, is not the ratio L, but rather $-2 * \log(L)$. This transformation via logarithm and multiplication results in a test statistic ($-2 \log \lambda$) with a well-known distribution: an asymptotic χ^2 distribution with 1 degree of freedom. This means that we can easily identify the associated probability value (p-value). If the result of the LLR test statistic is higher than or equal to 3.84, the null hypothesis is rejected (because it means $p < 0.05$) and the alternative hypothesis can be accepted. In which case, there is at least 95% certainty that the high relative frequency of the word in the specialised corpus is not due to chance. The asymptotic distribution is an approximation, which facilitates processing the problem of high numbers for word frequencies and corpus size. Keywords in the specialised corpus will be characterised by a high LLR value and a (very) low p-value. The higher the LLR value, the higher the keywords will be ranked on the typicality continuum.

Authors' addresses

Ann Bertels
Leuven Language Institute
KU Leuven
Dekenstraat 6 b 5302
B-3000, Leuven
Belgium

ann.bertels@ilt.kuleuven.be

Dirk Speelman
RU Quantitative Lexicology and Variational Linguistics
Faculty of Arts
KU Leuven
Blijde Inkomststraat 21 b 3308
B-3000, Leuven
Belgium

dirk.speelman@arts.kuleuven.be