A variational linguistics approach to term extraction.

Dirk De Hertog[1], Kris Heylen[1], Dirk Speelman[1], Hendrik Kockaert[2]

[1]University of Leuven, Belgium; [2]Lessius University College, Belgium

Automatic Term Extraction, Hypothesis-testing

Abstract

This paper describes how a toolset developed within variational linguistics for the purposes of identifying regional lexical variants, can be used in the field of term extraction. The notion of stable lexical marker analysis will be introduced as a method to quantify termhood as a function of both high relative frequency and uniform dispersion of single word units in a specialised domain. As such, the work is an extension of so called contrastive approaches to term extraction. The Belgian financial legal domain will serve as a case study and its results will be used to investigate how the method works and how it relates to approaches striving for the same goal.

1. Introduction

When the question is raised what automatic term extraction actually aims at, a straightforward answer would be that its goal is to extract the words typical for a domain.Term extraction literature makes the distinction between what is called termhood of a word and unithood. Termhood is defined as "the degree to which a stable lexical unit is related to some domain-specific concepts " (Wong, 2009). Unithood is "the degree to which a sequence of words is able to form a stable lexical unit" (Wong, 2009).

In the field of term extraction there have been many approaches that focus on the extraction of multiword units, and thus on the detection of unithood, assuming that multiword units comprise the majority of terms in most subject fields. This has the side-effect that mono-word term extraction has largely been disregarded. Lately some studies, such as Wong (2009) and Drouin (2008), have emerged that stress the importance of simple or mono-word extraction for both practical as theoretical reasons. From a practical viewpoint, not only is the prevalence

of multiword terms an insufficient reason to disregard mono-word terms, the exact ratio of mono- to multiword terms is hard to verify and might be domain dependent. From a theoretical stance the ISO-definition states that a term is “a verbal designation of a general concept in a specific subject field” (ISO 1087-1, 2000), and comprises both mono- and multi-word terms. As a term is seen as a conceptual unit, the extraction method should not a priori exclude one of these subsets.

While methods used to determine unithood, can function reasonably well using domain-internal frequency data, methods to determine termhood need more information to distinguish terms from non-terms, in the form of domain-external corpus evidence. These methods are called constrastive term extraction methods.

We would like to add that on top of comparatively higher frequencies, also consistency in use throughout the domain indicates a term's connection with that domain. Therefore we will present a method that calculates relative uniform dispersion as a part of TH.

2. State of the art

Contrastive approaches rely on the fact that terms are domain-specific, and as a consequence occur more frequently in their proper domain than they do in other domains. Several researchers have been using such contrastive approaches to determine TH.

The methods which are described shortly, can be split in two kinds of contrastive approaches. There are methods that use frequency or a transformation thereof to calculate TH. The approach is straightforward, easy to interpret and gives good results for recall. The other approaches uses some statistical test to see whether the expected frequency of a word, based on the distribution of that word in one of the corpora, is the same as the encountered frequency in the other corpus.

Tfidf is the oldest contrastive measure in use (e.g. Salton and Buckley (1988)) It measures the word's TH as a combination of its frequency and its inverted document frequency. It is a measure that originated in and is mostly used in an information retrieval settings context, to determine which words are good keywords for a given text. It increases the weight of less common words to make sure a query containing them, delivers the most relevant documents. Ahmad (2005) use a measure they refer to as the weirdness of a word, which is defined as the

result of the comparison of the word's normalised frequencies between an analysis corpus (AC) and a general language corpus, or reference corpus (RC). In this manner they "identify signatures of a specialism". Those words which combine high frequency and high weirdness are of most interest. Kit and Liu (2008) quantify the TH of a term candidate as its difference in frequency rank between a domain and a background corpus. This measure is based on the word's frequency for both types of corpora and is normalised by the total number of types in the corpus' vocabulary. In a second step they also enhance this value with information gathered from domain internal frequency. Chung (2003) uses a normalised frequency ratio to decide on TH. Wong (2007) proposes a similar technique that uses distributional behaviour of a word in opposing corpora to measure what he calls intra-domain distribution and cross-domain distributional behaviour. The first distribution is used to calculate a domain prevalence score, which measures the extent of the term's usage within the target domain. The second distribution is the basis for a domain tendency score, which measures the extent of term usage towards the target domain. Drouin (2008) compares precision and recall for the ranking of different measures used in hypothesis testing trying to determine which measure works best. Scott (1997) uses the χ^2 statistic to decide whether a word qualifies as a keyword, which he defines as" a word which occurs with unusual frequency in a given text."

Except for Wong (2007), the distribution of the terms across the domain is not investigated. The contrastive approaches treat the specialised and the general domain as a homogeneous and consistent whole, while this may in fact be a simplification of the material under investigation. The method investigated in this paper is also a contrastive approach, that tries to capture the term's consistency and distributional behaviour and has its origin in variational linguistics.

3. Stable Lexical Marker Analysis

Stable Lexical Marker Analysis was originally developed in the cross section between corpus linguistics (Kilgariff, 2001) and variational linguistics in the Labovian sociolinguistic tradition. The Method has been used to identify so-called lexical markers of different language varieties by Speelman, Gondelaers and Geeraerts (2006). An example is the difference in word use for the concept UNDERGROUND TRANSPORTATION NETWORK in American and British English. In this case, *subway* is said to be a lexical marker for American English and *underground* for British English. More specifically, the tool relies on

statistical hypothesis-testing of differences between word frequencies from different varieties. Important in the current context is that the specialised domains studied in terminology research can be considered as a specific language variety, or in terminological parlance *Language for Specific Purposes* (LSP), that is different from general language, a notion also expressed by Ahmad and Gillam (2005). The method developed for identifying lexical differences in two varieties, can thus be used for identifying terms using a specialised language corpus, the Analysis Corpus (AC), and a general language corpus, the Reference Corpus (RC). Being a lexical marker for a specialised corpus, can be seen as one of the necessary characteristics for qualifying as a term. The tool does not limit itself to a straightforward comparison between both corpora, as a keyword analysis based on a single hypothesis test would do (Scott 1997). It also calculates the dispersion of a word in a variety-specific corpus.

Stable lexical marker analysis defines the dispersion of a word as its consistency and stability within the domain and calculates this by using a pairwise comparison of a subdivision of both the RC and the AC. For example, both the specialized corpus (S) and the reference corpus (R) might be divided into 8 parts: {S1, S2, ... S8} and {R1,R2,... R8}. The next step is a pairwise comparison between all of the S-members and all of the R-members: {S1, R1} , {S1, R2} , ... {S8, R8}. In each pairwise comparison, statistical hypothesis-testing (e.g. a likelihoodratio-test) determines which words are lexical markers. A scoring scheme is applied so that a word gets credit for each pairwise comparison in which it is a lexical marker. If a word obtains a high score over all pairwise comparisons, it is called a stable lexical marker. For the example above, there are 64 possible combinations between group S and group R so the maximum score is 64 and the minimum score is 0. This way, the analysis provides a ranking that assigns the highest scores to the words that most consistently occur with a significant AC-RC frequency difference. In sum, the lexical stable marker method takes into account two properties of variety-specific lexical items. As other contrastive approaches, it extracts words that have an above-average frequency in the specialised corpus, but additionally, the method assures that these words have a high dispersion in the specialised corpus. This has the advantage of filtering out any frequency bias that might be introduced by just a part of the corpus. Such a locally clustered frequency bias is often caused by topical bias, as for instance introduced by a text that extensively discusses a topic otherwise unrelated to the domain.

4. Research Questions

The goal of this paper is to investigate how the SLMA-method, and more generally, the use of hypothesis-testing, along with its included measure for domain-consistency is of use in the field of term extraction. Does our method capture different information than other similar methods do? And if so, do we improve on these results? Firstly the method's ability to capture consistency of word use will be looked at. Secondly, an investigation into overlap with other measures, such as base frequency, as well as a short overview of precision and recall results, will provide evidence to which degree our method differs from these measures and captures different information.

As a case study, the Stable Lexical Marker method will be applied to mono-word term extraction from Dutch texts in the Belgian financial legal domain. Although the variational linguistic notion of lexical marker of a language variety does not completely overlap with the notion of a term in terminology research, it might well be fruitful to apply the method for the legal domain, as the domain is characterised by a very specific linguistic style that goes beyond the presence of terminological units in a strict sense, referring to clearly delineated concepts, but also involves rhetorical expressions and idiomatic language use. When these are considered as terminologically relevant LSP-characteristics - and we think they should - a sociolinguistically motivated analysis method like Stable Lexical Marker Analysis, might be better suited for terminology extraction than traditional term extraction methods.

5. Data collection and setup

The specialised corpus we have at our disposal is a financial legal corpus, obtained from EURLex[1] by collecting all documents with the EUROVOC keyword *finances*. It consists of material as diverse as reports, ordinances, decrees, written demands and notes totaling to a little over 27 million words. As a Dutch general language corpus we have material from five different national Belgian newspapers from the period 1999-2005 totaling to approximately 1.3 billion words. No linguistic preprocessing such as lemmatising or parsing has been done, firstly because the Dutch lemmatiser/parser at our disposal is known to generate higher error rates on legal texts, and secondly, the idiomatic and formulaic nature of legal language

[1] European Union law, see http://eur-lex.europa.eu

implies that, at least for some terms, terminological status is associated to word forms rather than lemmas.
For the SLMA, the RC has been randomly sampled and subdivided in 24 parts, breaking it down to slices of about 50 million words each. The AC has been divided into 4 parts, breaking it down to samples of a little under 7 million each. The size of the available corpora makes it possible to maintain a high enough frequency for the analysis of salient words. Frequency information for all word forms is taken into account and no filter whatsoever has been applied.

The SLMA uses Log Likelihood (G^2), calculated by R's built-in function, as its base statistic to measure disparity between observed and expected frequencies and to calculate p-values. For comparisons with low cell counts for which G^2 is not appropriate, , Fisher's exact test provides the p-values. The hypothesis-test itself was set at a p-value of .01 assuring the test is not too lenient towards small frequency differences with insufficient proof. The same settings for the p-value are used for a keyword analysis in which AC and RC are compared as a whole, i.e. without subdivision.

For the Rank Difference (RD) method all words in the AC and RC are first sorted by frequency, then alphabetically and are then given a rank. Unlike the method's inventors, we decided to include all AC-words, also those that were not found in the RC. Each word's AC and RC rank is normalised through division by the highest rank in the respective corpus. The difference of both normalised ranks (AC-Rank minus RC-rank) results in a RD-score that lies between 1 and -1, where the words closest to 1 are most marked for the AC. In a second step this RD-score was combined with AC-frequency, by a simple muliplication. We will refer to this method as the Frequency Adjusted Rank Difference (FA-RD) method. For a full description of the methods, see Kit and Liu (2008).

6. Results

6.1. General Overview

The raw frequency count file of the AC shows there are 179910 different words with a frequency higher than 1, and 86091 of those have a frequency above 5.

The RD method of Kit and Liu (2008) resulted in a list in which all words in the corpus had a continuous value for TH between -1 and 1. There are 219182 out of a total of 314384 words

with a positive score, but because of the continuous nature there is no single cut-off for termhood and in the discussion below we will take into account the full range of positive RD scores. FA-RD reranked the list obtained by RD, but in itself did not alter negative to positive scores or vice versa.

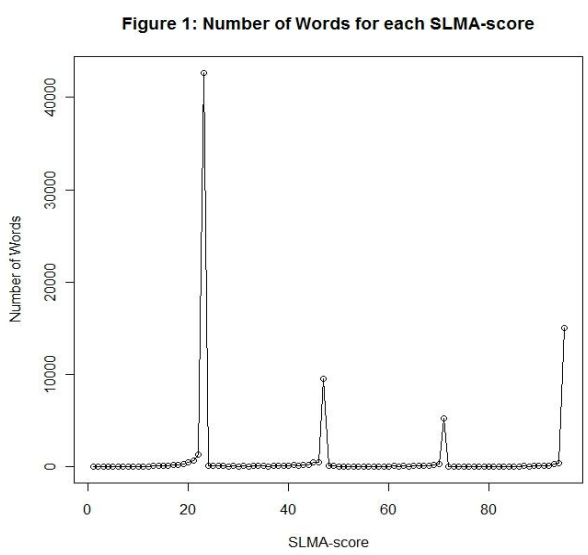
The key-word analysis with its single hypothesis-test reveals that 109,100 out of 179,910 words distribute differently across the two corpora. This number is quite high and it shows the corpora differ to a great degree in terms of word use. For the investigation into terms, this highly sensitive measure has an obvious limitation, viz. overgeneration. Additionally, the hypothesis-test proves that the word's frequency distribution is different for the two corpora, but it does not reveal in itself the size of this effect. Because of the binary nature of this significance test, by choosing a threshold for the p-values to decide on relevance, the information concerning the variation in likelihood of this word being a term is lost.

As stated abovethe hypothesis-test in key-word analysis uses log likelihood (G^2) as a divergence-from-expected measure . Interestingly, G^2 can also be used as a continuous measure for ranking term candidates (e.g. Drouin, 2008) and does not suffer from the binary decision of the hypothesis-test. The G^2 score corresponding to a p-value of .01 is 6.64, but as with RD, there is no a-priori cut-off and we will take into account the full range of G^2 scores.

The contribution of the SLMA method presented in this paper is that it starts from a hypothesis-test, like the keyword-analysis but it is nuanced by including a measure of dispersion. The division of the different corpora into 24 parts for the RC and 4 for the AC results in a maximum SLMA-score of 96. The higher a word's SLMA-score, the more consistent its frequency is significantly higher in the AC compared to the RC. 90,068 out of 179,910 words have a positive SLMA score. 15017 words have the highest score of 96. Note that although SLMA, like RD and G^2, can take on a whole range of values, it is more discrete in the sense that its values are necessarily integers.

6.2. Overlap

A first comparison between the keyword analysis and SLMA shows that 85,798 words out of a total of 90,068 words (95%) with a positive SLMA-score, are also found among the 109,100 significant term candidates according to the keyword analysis. Although this large overlap suggests that both methods capture the same information, a closer look at the break up of the SLMA scores gives a much more nuanced picture.



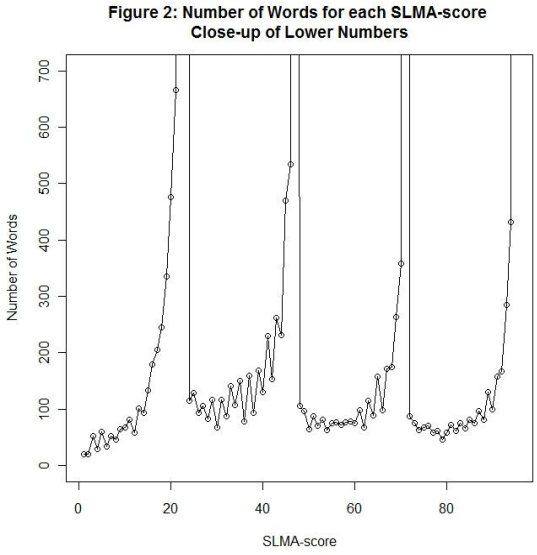


Figure 1 shows that the term candidates are unevenly distributed over the SLMA scores. This indicates that a lot of term candidates are not consistently more frequent throughout the specialised corpus as they do not pass all the pairwise significance tests of the SLMA. Moreover, the distribution of SLMA scores shows a clear peaked behaviour: the four local maxima at 24, 48, 72 and 96 reflect the subdivision of the AC in four parts. In other words, term candidates with these scores have a significant higher frequency in respectively 1, 2, 3 or all of the subparts of the AC. Aggregated frequency over the whole AC alone can obscure this lack of consistency: Table 1 shows that similar aggregated frequencies can indeed have different SLMA-scores.

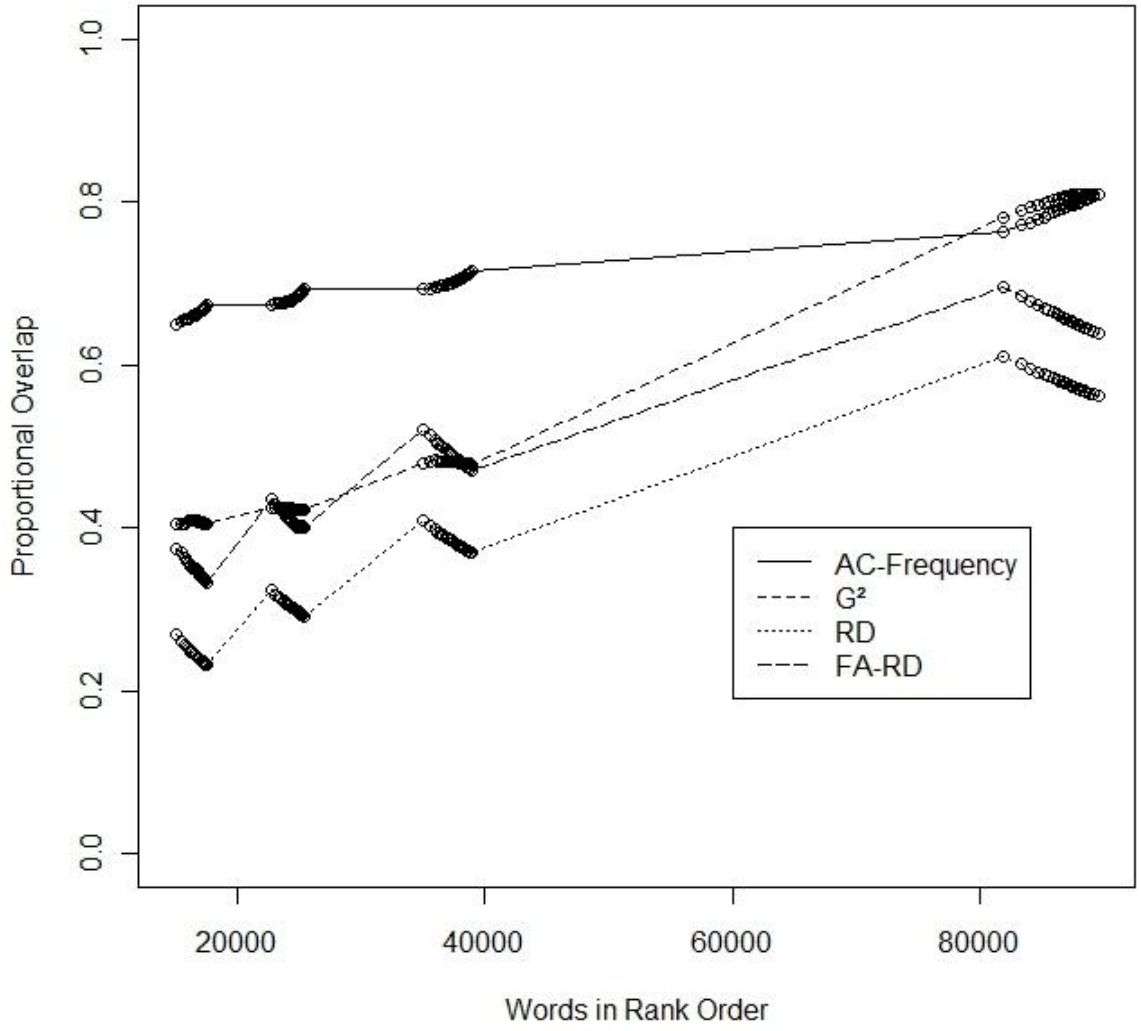
Figure 2 zooms in on the lower frequency ranges and shows that the variation in SLMA-scores between the peaks is mainly due to the RC-subparts: These words are not frequent enough in a given subpart of the AC to pass all pair-wise tests with the 24 RC-subparts. In summary, the SLMA-scoring, diversifies the words in terms of their consistency of relative high frequency, and the main diversification is influenced by the data gathered from the AC and only to a lesser degree by that from the RC.

Table 1: Lower Local Maxima SLMA-scoring Words with AC- and RC-frequency			
	SLMA-score	AC-frequency	RC-frequency
tinverbindingen "tin alloys"	24	29	1
uitlaatemissies "exhaust pipe emissions"	24	34	1
disadvantaged	24	34	1
perpetuals	24	24	2
telecommunicatiegebied "telecommunications area"	24	11	1
icelandic	48	127	1
kandidaat-verwerver "candidate recruiter"	48	123	1
null	48	939	8
Ajinomoto (Japanese company)	48	209	2
Ryanair (Belgian airport company)	48	717	7

Let us now turn to the comparison of SLMA with the methods that also provide a ranking of term candidates, rather than just a binary division. Figure 3 shows the overlap of term-candidates (in percentage on the Y-axis) between SLMA and one of the other methods, given that all words up to specific rank are taken into account (ranks are sorted from highly ranked to lowly ranked on the X-axis). Note that the SLMA has a lot of ties (equally ranked term candidates) because it assigns discrete scores only (integers from 0 to 96). Therefore, overlap has only been calculated at the ranks corresponding to these discrete scores, which can be identified in the figure by the small circles[2]. The lines between circles are smoothed fits. The overlap percentages converge naturally towards 1 when all words are taken into account.

[2] Because of the high number of ties, rank correlations of SLMA-score with other measures are unreliable and not calculated.

Figure 3: Proportional Overlap of SLMA-score with other Measures



The plot shows that the SLMA ranking has the highest overlap with raw frequency ranks in the specialised corpus. In other words SLMA-scores are strongly influenced by AC-frequency: 65% of the 15000 highest scoring words for SLMA also belong to the 15000 most frequent words in the AC. This doesn't come as a surprise, since a relative high frequency in the AC is one of the elements contributing to a high SLMA score. Yet, we also see that SLMA score is not fully determined by AC frequency. A closer look at the most frequent AC-words reveals that SLMA successfully removes a number of non- terminological general language words: the non-overlapping words in the top 100 most frequent words are elements, such as {*een* (a), *dat* (that), *is* (is), *zijn* (are), *aan* (particle - to), *om* (to), *niet* (not), *kan* (can), *dan* (than) , *uit* (particle - out), *als* (if), *hebben* (have), *er* (it, there), *naar* (to), *ook* (also), *geen* (no, none), *meer* (more) , *hun* (theirs)}. However, it cannot be said that the method filters out

all general language elements because elements like {*de* (the), *van* (of), *het* (the), *en* (and), *in* (in), *voor* (before), *op* (on) , *te* (to), *met* (with), *die* (that), *worden* (become)} are also among the top 100 words with the highest SLMA-scores. This shows that general language elements with a very high AC-frequency are not pushed down in rank.

The second highest overlap is between SLMA and G^2. Since G^2 is the underlying statistic used in the hypothesis tests reflected in the SLMA-scores, some similarity between the methods is to be expected, but with a figure of just 40% for the 15000 highest ranking words, this overlap is relatively low. However, remember that the G^2 ranking shown is obtained by comparing the AC and RC in their entirety, whereas SLMA subdivides the corpora. In other words, the consistency checking of SLMA shifts the rank order considerably.

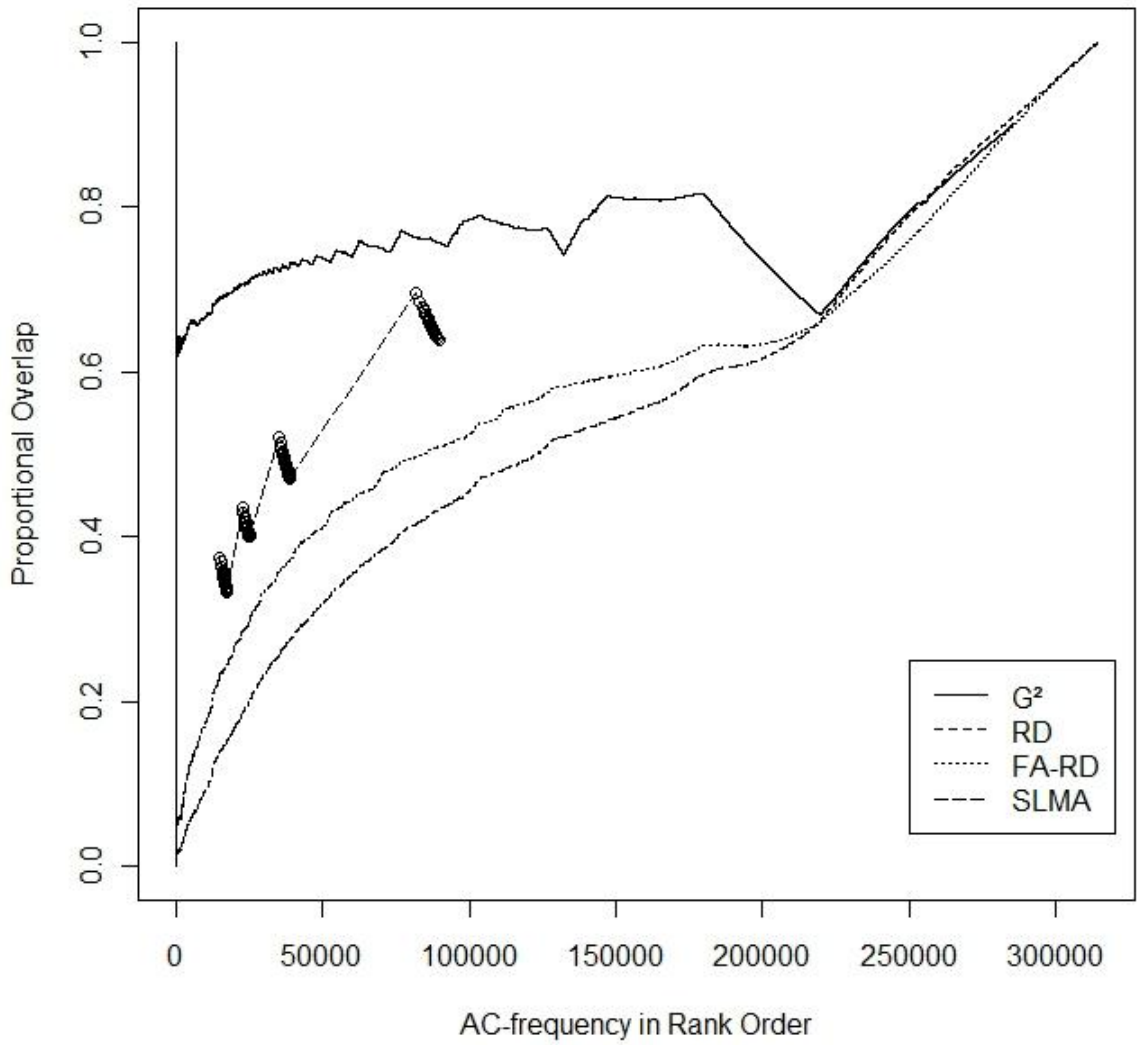
The overlap with RD and FA-RD is lower than with the other measures. Especially in the section with top SLMA-scores, a low overlap can be seen. Clearly, RD captures different information than SLMA, although the AC-frequency adjustment of FA-RD brings both methods closer together again, which is not surprising given the higher overlap between AC-frequency ranking and SLMA-ranking. Interestingly, the same peaked behaviour as was seen in the separate discussion of SLMA-scores surfaces again. Apparently, overlap goes up when the SLMA-ranking corresponds to high-frequencies in one subpart of the AC, but drops again if SLMA-ranks correspond to non-consistent higher frequencies relative to the RC. Because SLMA and RD are the two methods that are most specifically geared towards term extraction, we turn to a few examples to get a better idea of their differences. The top and bottom section of Table 2 highlights some of the words that are ranked significantly differently by both methods. The Lithuanian words found at the lower portion of table 2 occur significantly often in but one subpart of the corpus, showing that some of the documents in the AC were drafted in Lithuanian. Our method successfully singles out these words as being irrelevant for the Dutch legal domain. The top section of this table are all words that are clearly situated within the legal financial domain. RD reranks these words in such a way that they are moved towards neutrality, because the word's frequency in the RC is higher than average. SLMA captures their consistent use in the AC and its importance in the domain. All these words have middle-frequencies that show an even distribution in both corpora. Sometimes however, SLMA wrongly classifies a word as a term whereas RD classifies it correctly as neutral (e.g. *zinvol* 'meaningful'). The frequency per million tells us it concerns an extremely frequent word, which hints at the fact that the SLMA-method might not be suited for words with this

frequency profile, whereas RD is. For lower-frequency words (e.g. *vergissingen* 'mistakes') SLMA does not suffer from this overgeneration.

Table 2: Overview of Words with SLMA score, AC- and RC-frequency, RD-rank and RD-score

	SLMA-score	freq AC (/Million)	freq RC (/Million)	RD-Rank	RD-score
BTW-plichtige "liable toVAT"	94	1,852	0,061	220769	-0,003
renteloos "Interest free"	94	1,000	0,102	239142	-0,060
jaartotaal "annual total"	92	0,963	0,080	239046	-0,066
substituten "substitutes"	77	2,259	0,392	233816	-0,044
hervormingspakket "reform package"	68	0,741	0,052	240416	-0,073
schuldvergelijking "debt equation"	68	1,148	0,008	214123	0,014
verzekeraar "insurer"	96	29,148	10,388	223179	-0,009
solvabiliteit "solvability"	96	16,926	0,824	223184	-0,009
zinvol "meaningful"	90	10,333	5,172	226731	-0,010
vergissingen "mistakes"	33	2,556	1,415	234118	-0,045
opmerkzaam "observant"	33	0,630	0,138	245318	-0,098
koelsystemen "cooling systems"	13	0,407	0,198	252042	-0,136
meteorologische "meteorological"	12	0,889	0,463	243030	-0,086
gunsten "favours"	27	2,556	0,828	233756	-0,043
eg-verdrag "EC-treaty"	24	8,781	0,026	9	0,998
asignavimai "appropriations" (Lithuanian)	24	3,289	0,000	30	0,996
asignavimas "appropriation" (Lithuanian)	24	1,475	0,000	74	0,992
išmoka "allowance" (Lithuanian)	24	1,384	0,000	83	0,992
patvirtinimas "confirmation" (Lithuanian)	24	1,383	0,000	84	0,992

Figure 4: Proportional Overlap of AC-frequency with Other Measures



Finally, Figure 4 looks at the relation between raw AC-frequency ranking, as a sort of natural baseline, and the ranking by the G^2 and RD methods. The overlap with SLMA is repeated for ease of reference. Like SLMA, G^2 appears to be also highly influenced by base frequency, which helps explain the relative higher overlap between SLMA and G^2. The opposite is true for RD: it shows less frequency bias, which partially explains its lower overlap with SLMA.

6.3. Recall and precision

In the previous section, we established that the different methods under investigation rank term candidates differently, at least to a certain extent. In this section, we will explore which methods are better at the task of term extraction for a known set of terms. As a reference list, a collection of single words was brought together from Moor's legal dictionary (CD-ROM, 2006) and legal terms from a spelling list (MS-WORD specialised word list[3]). Because no sizable specialised reference list for *financial* legal terms was found, these general legal terms are used as an approximation of the domain's language use. Only word forms consisting only of letters (as opposed to digits) have been taken into consideration.

Before the results are discussed some remarks are in order about the nature of the reference list used to calculate precision and recall. The sources from which the list is compiled contain a lot of words that also occur in general language. This might reflect the special relationship between general language and legal language when compared to other LSP's but it has two consequences for the results. On the one hand they provide an underestimate because this rather general legal dictionary is not exhaustive in its coverage, while on the other hand it provides an overestimate because general language elements are also included in the reference list. This makes it difficult to provide an accurate measure for performance, especially for SLMA, as theoretically its key strength is filtering the general language elements out. A second remark concerns the nature of the texts that are used in the financial law corpus. The legal domain is characterized by a division in primary texts (the law texts themselves) and secondary texts (discussions and interpretations of the law texts). Our corpus mainly contains primary texts and some terms typically used in secondary texts are likely to be missing.

[3] http://www.microsoft.com/downloads/details.aspx?FamilyID=159e1f83-804f-4e28-ba47-7d4bd3715f5f&DisplayLang=nl

Figure 5: Recall

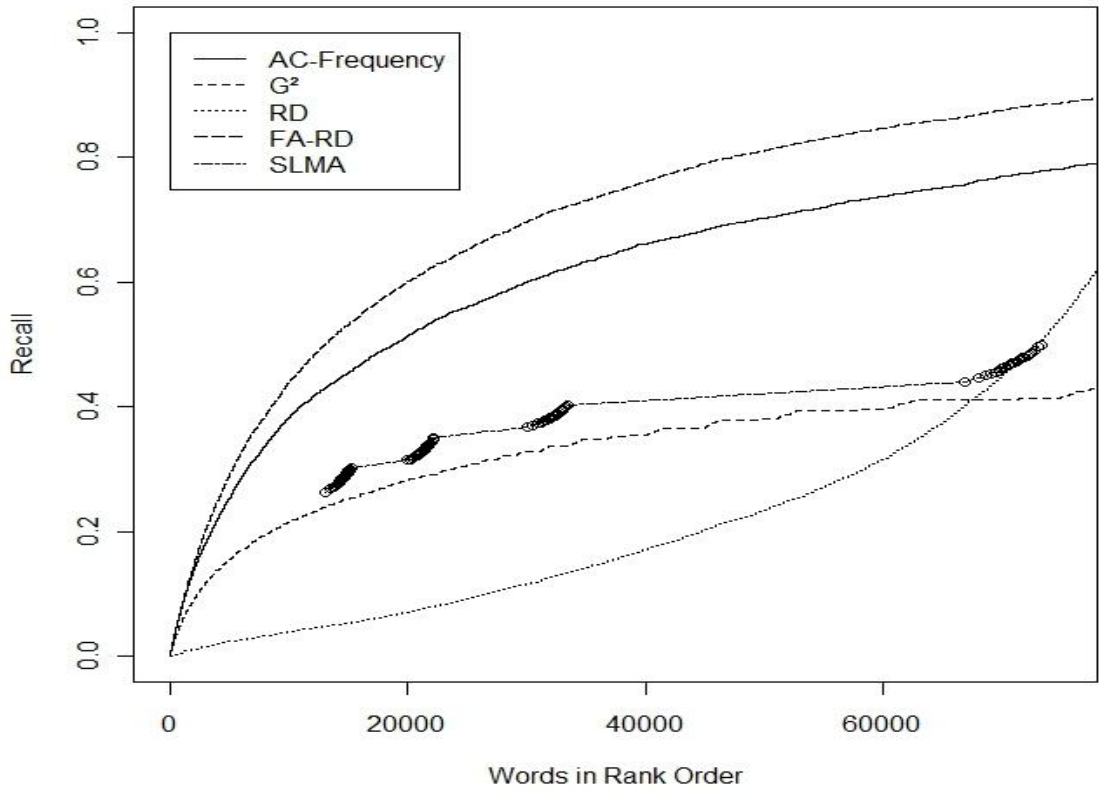


Figure 6: Precision

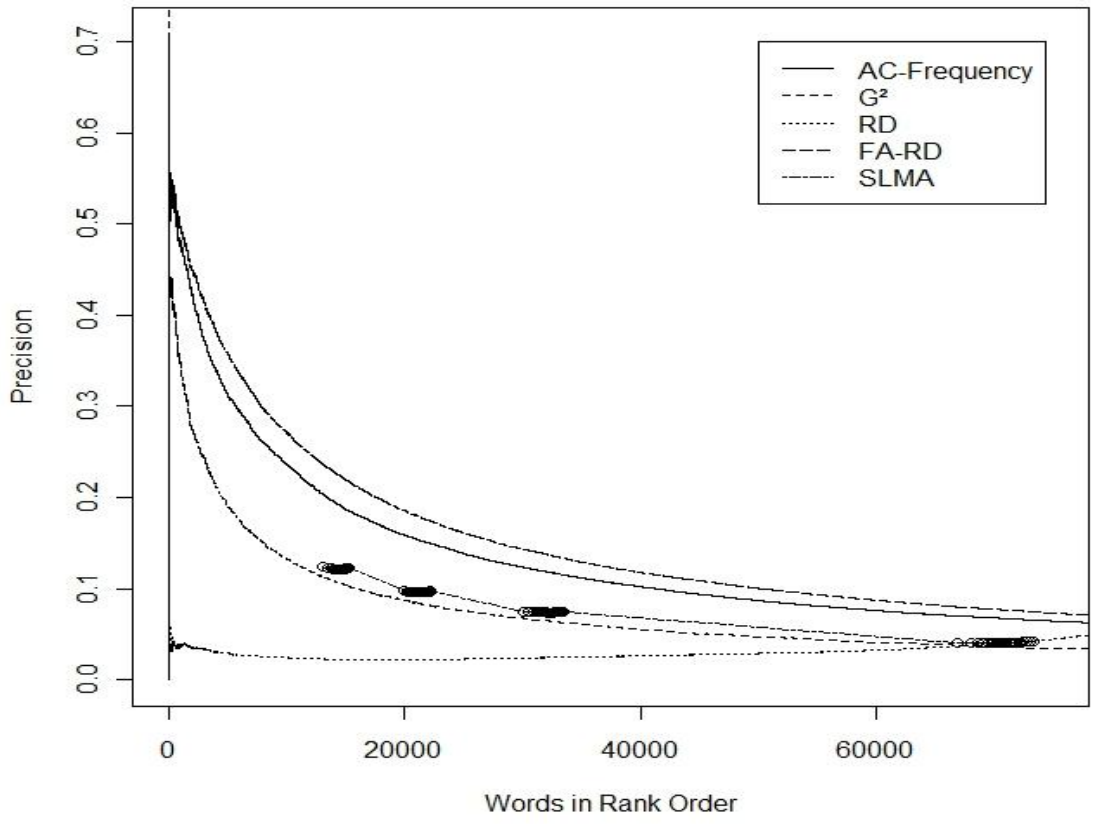


Figure 5 shows the cumulative recall plot up to the 75,000th ranked word for the previously discussed measure. Figure 6 deals with the precision rates for these 75,000 highest ranked words. Note again that the SLMA-score takes on a discrete range values and that only ranks corresponding to these discrete values are plotted. As we can see in figures 5 and 6 the method that captures most terms and also achieves the highest precision rate (albeit still low in absolute figures), is the frequency adjusted RD method from Liu and Kit. Raw AC-frequency is a close second and greatly influences the results obtained from the former method. Since the RD-method in itself clearly underachieves both in terms of precision and recall, the main success component of FA-RD is the raw AC-frequency, rather than the rank difference measure. SLMA comes in third, suggesting that, at least for this reference list, its hypothesis testing paradigm is wrong in downplaying the importance of domain-internal frequency. On the other hand SLMA is still slightly better than G^2, indicating that, within a paradigm based on measuring the divergence between observed and expected frequencies, SLMA's consistency checking mechanism with multiple comparisons, rather than just one, might be useful.

7. General discussion

As the two methods specifically designed for Term Extraction, we will concentrate here on how RD and SLMA relate to each other. The results from the overlap section show that the AC-frequency to a great degree influence the term-candidates retrieved by SLMA in the sense that a high AC's frequency-count is a determining factor of acquiring high SLMA-scores. At the same time and paradoxically, the precision and recall plots suggest that this AC-frequency information is not exploited enough. RD's dissociating from AC-frequency and its resulting poor performance stresses even more the importance of domain internal frequency to gauge a word's chances of being a term. Consequently, Kit and Liu are right to incorporate raw domain-internal frequency to augment the results of their method. Yet, if domain-internal frequency alone captures all information, automatic term extraction would no longer be a research topic. After all, there are still general language elements that are also commonly used in the domain's language and it is exactly the task of a term extraction method to filter out these elements from the top section of the term candidate list. However, the problem of hypothesis testing methods like SLMA is that there are different classes of words that fall under the category of general language elements. On the one hand there are the general

language elements that refer to every-day concepts, and which occur moderately often in any text. On the other hand there are the function words of a language which occur pervasively and with high frequency whenever the language is used. It seems hypothesis tests are not good at handling this second type of words. As a general note it has to be said that G^2, χ^2or any other measure that captures divergence between expected and observed frequencies and uses this for hypothesis testing, is known to exhibit a sensitivity both to extremely high and extremely low frequencies. For words with high expected frequencies even relatively small differences will be treated as significant. Low frequencies cause the statistics to be unreliable with regard to the decision whether the encountered frequency differences are systematic. This explains why the method overgenerates for extremely high frequencies, such as the pervasive class of function words, or such as some very popular concepts of the first class. RD on the other hand does seem to manage these unwanted high frequency words rather well. For words in the middle frequency range, SLMA succeeds in singling out words with a high consistency and an overall reasonably high frequency. These words are incorrectly downgraded by RD. Words with an extreme frequency difference between AC and RC are recognised by both methods. While low frequency words prove problematic to decide on TH for either method.

To summarise, each method has its strength based on what we will refer to as the frequency profile of the words under investigation. We still think that it is not just AC-frequency and RC-frequency that are important in determining TH, but also information on dispersion, or consistency in word use throughout a domain. From the overlap plot between high SLMA-score and RD it is clear that both methods capture different information. The investigation into which words were singled out made clear that both methods exhibit sensitivities with regard to information the other method processes correctly.

8. Conclusions and future work

It has been shown that contrastive approaches rerank the list of candidate terms in such a way that general language element words are pushed more towards neutrality. However Hypothesis-testing as such is sensitive to high frequent words, making it only partially successful in this endeavour. It has become clear that the most determining factor of the SLMA-scoring distribution are in fact the AC-frequency counts. Recall measures show that raw frequency is a good guiding factor for including terms, so a slight bias towards frequency

should not prove problematic. Some improvements are necessary however in order to exclude false positives. The hypothesis-testing, even when incorporating some measure of dispersion will benefit from including some measure that captures effect size. The results prove that each method decides differently on TH depending on the word's frequency profile and that an motivated decision based on this frequency profile can be made as to which method to confide in.

Some concerns regarding methodological choices will have to be further investigated in future work. By subdividing the AC, we are aware that we might introduce data sparseness. As such, words with an overall low-frequency in the AC cannot be straightforwardly discarded as a non-term based on the acquisition of a low SLMA-score alone. For this reason the relationship between frequency and SLMA-score needs further attention. Another remark concerns the choice of corpus subdivision. Because this division influenced the SLMA-score, differences in setup of the corpora, such as size, and number of subdivisions, will have to be examined more.

Furthermore, future work will include all forms of term normalisation to resolve the word forms to conceptual units, such as lemmatisation, multi-word unit detection, and the detection of term variants, to improve on the precision of base frequency counts.

References

Chung, T. 2003. 'A corpus comparison approach for terminology extraction.' *Terminology* 9(2), 221-246.

Dunning, T. 1993. 'Accurate methods for the statistics of surprise and coincidence.' *Computational Linguistics* 19(1), 61-74.

Drouin, P. and Doll, F. 'Quantifying TH through Corpus Comparison' In Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds) : *Managing Ontologies and Lexical Resources*, 191-206. Copenhagen, 2008.

Gillam, L. and Ahmad, K. 2005. 'Pattern Mining Across Domain-Specific Text Collections.' In *Lecture Notes in Computer Science* 3587(2005). Springer Berlin-Heidelberg. 570-579.

ISO 1087-1 (2000) E/F. International Organisation for Standardisation

Kilgarriff, A. 2001.'Comparing corpora.' *International Journal of Corpus Linguistics* 6(1), 97-133.

Kit, C and Liu, X. 2008. 'Measuring mono-word TH by RD via corpus comparison.' *Terminology* 14(2), 204-229.

Moors, J. 'Juridisch woordenboek op CD-ROM (N-F/F-N)'. 2006.

Pedersen, T. 1996. "Fishing for exactness." In *Proceedings of the South-Central SAS Users Group Conference*, Texas.

Salton, G. and Buckley, C. 1988. 'Term weighting approaches in automatic text retrieval.' *Information processing and management* 24(5), 513-523.

Scott, M. (1997. 'Pc analysis of key words - and key key words.' *System*, 25, 233-245.

Speelman, D., Gondelaers, S. and Geeraerts, D. 2006. 'A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch.' In Andrew Wilson, Dawn Archer & Paul Rayson (red.), *Corpus Linguistics around the World* 195-202. Amsterdam: Rodopi.

Wong, W., Liu, W., Bennamoun, M. (2007). 'Determining TH for learning domain ontologies using domain prevalence and tendency.' In *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, Vol. 70. Gold Coast, Australia, 2007).