# The prevalence of multiword term candidates in a legal corpus

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

[1] QLVL, KU Leuven (University of Leuven), Belgium
{dirk.dehertog,kris.heylen,dirk.speelman}@arts.kuleuven.be
[2] Lessius University College, Belgium
hendrik.kockaert@lessius.eu

**Abstract.** Many approaches to term extraction focus on the extraction of multiword units, assuming that multiword units comprise the majority of terms in most subject fields. However, this supposed prevalence of multiword terms has gone largely untested in the literature. In this paper, we perform a quantitative corpus-based analysis of the claim that multiword units are more technical than single word units, and that multiword units are more widespread in specialized domains. As a case study, we look at Dutch terminology from the Belgian legal domain. First, the relevant units are extracted using linguistic filters and an algorithm to identify Dutch compounds and multiword units. In a second step, we calculate for all units an association measure that captures the degree to which a linguistic unit belongs to the domain. Thirdly, we analyze the relationship between the units' technicality, frequency and their status as a simplex, compound or multiword unit.

**Keywords:** Legal Terminology, Automatic Term Extraction, Multi Word Units

## 1 Introduction

Many early approaches to term extraction focus on the extraction of multiword units [1, 2]. These studies often assume that multiword units com-prise the majority of terms in most subject fields. Daille [3] uses an existing term base to determine the amount of multiword terms found for a technical domain and confirms this assumption. However, there are two topics in the field of term extraction that increasingly received attention by the research community and that go beyond the traditional focus on multiword terms. A first development coming into focus from the late nineties on is the inclusion of domain external frequencies on top of domain internal information in the form of contrastive extraction methods [4, 5]. A second development is the application of term extraction beyond purely technical domains, such as the legal domain.

Contrastive methods come in many flavours, and were first investigated for information extraction purposes and keyword detection. The meth-ods employed in these contrastive approaches range from contrasting frequency signatures of

Aguado de Cea et al. (Eds.): Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), pp.283-290. 19-22 June 2012, Madrid, Spain

words, to the use of more advanced statistics and heuristics. The goal of these methods remains the same throughout, namely detect those words that are most characteristic for a conglomerate of texts. The use of two distinct corpora to assess the association of lexical items to a specific corpus, opens up the possibility to distinguish general language elements from more specialized ones. Whereas domain internal frequencies of single words are not very informative (e.g. they cannot distinguish highly frequent terms from mere function words), the use of contrastive frequency information does make it possible to assess the termhood of single word units. As the availability of larger corpora is ever increasing, data sparseness is becoming less of an issue and even the investigation of multiword units using a contrastive approach has become feasible. In the context of term extraction the assumption is made that these methods allow for an automatic detection of terms because characteristic words of any given text are more likely to be items of interest for those professionals coming into contact with the specialised domain. The application of term extraction beyond purely technical domains, such as the legal domain is a second topic that has seen an increase in re-searchers' interest. The nature of the terminological unit for the legal domain, i.e. the definition of a term, is currently under discussion [7]. While term extraction for purely technical domains concentrates itself mostly on the noun phrase, as the description of concrete objects is central to the understanding of technical texts, legalese is characterised not only by the use of specific objects. Legal actions involve processes expressed by verbs, fixed expressions grant the texts a legal character, and the interaction be-tween processes and the subjects or objects to which they apply are often combined in a specific phraseology found exclusively in legal texts. To get a more precise linguistic definition of the sort of word combinations that occur in legal language, we can have a look at the types of phrasemes that Kjaer [8] distinguishes for German, which is linguistically closely related to Dutch.

1. multi-word terms: (Adjective + Noun) *hoofdstedelijke regering*
2. Latin multi-word terms: (loan words) *ex officio*
3. Fachwendungen (LSP phrases): (Noun + Verb) *een advies verlenen*
4. Funktionsverbgefge (Support Verb Construction): ((Preposition) + Noun + Verb) *een beslissing nemen*
5. (Archaic) Formulae: ($X + X^{+}$) *Hebben Wij besloten en besluiten Wij*

Because of the specific language use in legal discourse [9] the application of term extraction to the legal domain might warrant precaution with regard to assumptions made to the nature of its terminological units. As such, it is not only unsure whether multiword units make up the majority of extracted units in this domain, it is also unclear whether they actually are more technical than their single word counterparts. This paper will investigate these two questions.

A corpus based approach is used that takes into account the unit-level of the term candidates. The case study presented here limits itself to the investigation of noun phrases in Dutch. From a methodological point of view it can be seen as an example of how to detect and extract noun phrases from languages with similar compounding mechanisms, such as German.

From a Dutch linguistic perspective, we need to distinguish four groups of noun phrases. Firstly, there are simplex nouns consisting of only one word stem. Secondly, we have compound nouns, which consist of two noun stems, but unlike most English compounds, they are written as an orthographic unit in Dutch. Thirdly, there are the adjective noun combinations which are truly multiword terms in Dutch. Finally, there are noun preposition noun combinations and variations thereof, which will not be investigated further in this case study. Both compounds, adjective-noun combinations and noun preposition noun combinations can be considered to be complex units, and would be multiword units in many other languages.

**Table 1.** Examples of Dutch noun phrases and their English translational equivalents

| | Dutch | English |
|---|---|---|
| Simplex noun | procedure | procedure |
| Noun-noun compound | procedurefout | procedural infringement |
| Noun preposition noun combination | Raad van State | Council of State |
| Adjective noun combination | schriftelijke procedure | written procedure |

## 2 Methodology

### 2.1 Corpus based Approach

We have two corpora at our disposal that have undergone linguistic pre-processing in the form of POS-tagging. As a specialised corpus, we use a digital version of 'Het Belgisch Staatsblad', which contains official public announcements concerning legal matters, such as laws, royal decrees, decrees, ordinances, etc. The material spans the period of 1997 to 2006 and counts 80 million words. As a reference corpus of Dutch general language, we have a collection of Belgian Dutch-language newspaper materials spanning the years 1999-2005 which consists of a total of over 1.3 billion words.

### 2.2 Multiword Unit Identification

There are two distinct linguistic patterns for Dutch we will use in our investigation of complex terminological units. The first pattern are noun-noun compounds, the most productive compounding system for Dutch. Compounds have the advantage that we do not have to decide whether they are units or not, for they are orthographic units in Dutch. The second pattern are adjective-noun combinations. Both groups are the Dutch equivalent for what is referred to as multiword units in other languages such as French or English.

The compounding system for noun-noun combinations in Dutch is extremely productive. Theoretically, any two nouns can be combined using straightforward

combination rules: the two words are agglutinated with a binding particle. This binding particle comes in the form of -e, -en, -s or a hyphen. The detection of compounds thus requires some base vocabulary to test whether the compound can be broken down into two or more existing words, along with any of the possible binding particles. As a list of possible split-candidates a selection took place of all nouns from the 'Belgisch Staatsblad' with a minimum frequency of five, and all nouns in the newspaper material with a minimum frequency of twenty. The minimum string length of the split candidates was set at three. The frequency threshold and the minimum length criterion were chosen as to minimize the amount of non-existing 'garbage' words that found their way in the raw corpus material. Manual removal of faulty words from the base vocabulary took place to further improve the results. Because the compounding process is recursive in nature, theoretically multiple splits are possible. A heuristic that selects those split candidates that are most likely to exist as words in their own right proved to be the most adequate at selecting the correct candidates. For instance the word *basisloonberekening* (basic wage calculation) could be split in either *basis* and *loonberekening*, or *basisloon* and *berekening*. The second split is the one which is linguistically correct and this is reflected in the frequencies of the candidates in the two splits; 174283 and 134, versus 383 and 7178, . In this fashion 20700 noun-noun compounds have been detected.

The second linguistic group under investigation are adjective noun combinations. The POS-tagged corpus allows for a templatic extraction of all adjective noun combinations. Wrong tags have been manually removed from the candidate list. On top of this, all deictic adjectives (e.g. dit [this], vorige [previous], volgende [following]) and any numeric modifiers have been filtered out. A frequency threshold of four for the found collocations was set. The resulting list thus consists of rather frequent adjective noun collocations (35875).

### 2.3 Stable Marker Effect size Analysis

There are different association measures available to express the degree to which a word belongs to a domain-specific corpus. Different approaches use, amongst others, frequency information, statistical divergence measures such as $\chi^2$ or log-likelihood, information theoretic measures such as Mutual Information, or other heuristics. We opt for the method of Stable Marker Effect size Analysis (SMEA), a further refinement of the Stable Lexical Marker Analysis (SLMA) [10]. It measures the association of a word to a domain-specific corpus rather than a general language corpus and takes into account three key pieces of information:

1. A statistical hypothesis test (log-likelihood ratio) for determining a cut-off for term candidates
2. A gradual effect size measure for the strength of association (odds ra-tio) to sort term candidates
3. A procedure to check the consistency with which term candidates have an above-average domain-specific frequency

SLMA was developed in the cross section between corpus linguistics [11] and variational linguistics in the tradition of Labov [12], and is used to identify so called lexical markers of different language varieties [13]. It is conceptually based on the keyword-analysis introduced by Scott [14]. A keyword analysis uses frequency information of a word from two different corpora to assess whether a word is associated to one of them. A straightforward comparison between two corpora, based on traditional keyword analysis [14] suffers from topical bias however. If a small part of the corpus deals with a specific topic, the words associated with that topic could have an above average frequency, while the dispersion of those words in the entire corpus shows that they are generally not associated to the corpus as a whole. This is why the marker analysis score is calculated specifically to reflect the dispersion of a word, and hence the consistency and stability of its difference in usage between two corpora. To make it more concrete: two corpora (A and B), each of which is representative for a domain under investigation, might be divided into 8 parts: $\{A1, A2, ...A8\}$ and $\{B1, B2, ...B8\}$. The next step is a pairwise comparison between all of the A-members and all of the B-members: $\{A1, B1\}$ , $\{A1, B2\}$ , ... $\{A8, B8\}$. In each pairwise comparison, statistical hypothesis tests determine which words are lexical markers that occur significantly more frequently in the A-corpus as compared to the B-corpus. P-values used for the hypothesis test underlying the method have been calculated using Fisher's exact test [15] for low-frequency words, for which the approximation of p-values of log-likelihood is not trustworthy, while log likelihood is used for words that have a high enough frequency. If a word is found to be associated to a certain part of the corpus in such a pairwise comparison, the averaged odds ratios are calculated to capture the odds to which a word is associated to a corpus. This approach permits to capture what is called effect size in statistical terms, the actual size of the strength of association. A further log transformation of the odds ratios improves the ease of interpretation of the results with regard to markedness. The scale after the log transformation ranges from high negative values to high positive values. Higher values mean stronger association. Evidently, negative association to one corpus implies positive association to the other one. The formula to obtain the score is:

$$SMEA(w_k, A, B) = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} (log(\frac{F_{w_k}^{A^i} / F_{\neg w_k}^{A^i}}{F_{w_k}^{B^j} / F_{\neg w_k}^{B^j}}) * S(F_{w_k}^{A^i}, F_{\neg w_k}^{A^i}, F_{w_k}^{B^j}, F_{\neg w_k}^{B^j}))$$

For a given word $w_k$, the log odds ratio is calculated by taking into account the word's frequency in the $i^{th}$ partition in corpus A and the frequency of all other words that are not $w_k$. This is repeated for the occurrence of the word in corpus B. $S()$ is a boolean function with value 1 if the frequency distribution of the word is significantly different in corpus A when compared to corpus B, and 0 otherwise. The sum of the values is then divided by the total number of comparisons.

For the current research question we consider the SMEA-score to reflect the technicality of the units under investigation. We are aware that this score only partially reflects an expert's judgement on termhood of the units, but consider

it a good approximation for termhood given the lack of resources needed for manual evaluation.

## 3 Results

The barplot in figure 1 visually presents the results of the investigation of the distribution of the three types noun phrases. It shows the association score of the different NP types to the specialized corpus, as calculated with the Stable Lexical Marker Analysis method. Table 2 is a summary of the information in the barplot.

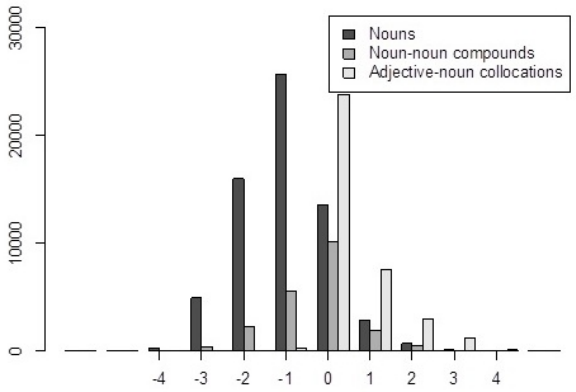


**Fig. 1.** SMEA-score of nouns, noun-noun compounds, and adjective noun collocations by number of occurrences

With a total of 55.134 simplex nouns, this group is slightly smaller than the combined occurrences of noun-noun compounds (20.700) and adjective-noun collocations (35875). The corpus thus contains more complex units than it does simplex nouns. This picture becomes even more outspoken if we look at the noun phrases that have a clear positive association to the specialized corpus and that can be considered to be term candidates. Among noun phrases with an SLMA score of 1 or higher, 66.8% (11802) are adjective-noun pairs, 13.6% (2406) are compounds and merely 19.5% (3441) are simplex nouns.

The distribution of both the simplex noun group and the noun-noun compounds show a Gaussian distribution with regard to association to the specialized corpus. A peak of number of occurrences is seen at the neutral 0-value. The adjective-noun collocations show a right skewed distribution, starting with its peak at the 0-value. Noun-noun compounds thus behave similar to the simplices in the Dutch legal domain, with an even distribution across the SLMA-scores.

**Table 2.** Overview of investigated units and their SMEA-scores

| | nouns | | noun-noun comp. | | noun-adj. coll. | | |
|---|---|---|---|---|---|---|---|
| SMEA | Absolute | Relative | Absolute | Relative | Absolute | Relative | Total |
| -5 | 4 | 80.00% | 1 | 20.00% | 0 | 0.00% | 5 |
| -4 | 194 | 91.51% | 18 | 8.49% | 0 | 0.00% | 212 |
| -3 | 2480 | 86.41% | 390 | 13.59% | 0 | 0.00% | 2870 |
| -2 | 9771 | 81.26% | 2220 | 18.46% | 34 | 0.28% | 12025 |
| -1 | 25678 | 81.55% | 5542 | 17.60% | 269 | 0.85% | 31489 |
| 0 | 13566 | 28.58% | 10123 | 21.33% | 23770 | 50.09% | 47459 |
| 1 | 2788 | 22.78% | 1911 | 15.61% | 7542 | 61.61% | 12241 |
| 2 | 585 | 14.51% | 464 | 11.51% | 2984 | 73.99% | 4033 |
| 3 | 67 | 5.15% | 31 | 2.38% | 1204 | 92.47% | 1302 |
| 4 | 1 | 1.39% | 0 | 0.00% | 71 | 98.61% | 72 |
| 5 | 0 | 0.00% | 0 | 0.00% | 1 | 100.00% | 1 |
| Total | 55134 | | 20700 | | 35875 | | 111709 |

They vary from high association with the specialized corpus to very negative association. The adjective-noun combinations are overall more associated to the specialized corpus.

This results in the confirmation of the initial research question that complex units are more prevalent and more associated to the legal corpus than their simplex counterparts, at least as far as the noun phrases for the Dutch legal domain are concerned. This result is mainly due to the behaviour of adjective-noun collocations, showing higher association towards the specialized corpus. However, simplices and compounds together, i.e. term candidates written as one orthographic unit, still represent a sizeable share of domain-specific words (1/3 of NPs with an SLMA score above 1) and cannot be neglected when performing term extraction for a compounding language like Dutch.

## References


1. Bourigault, D.: Surface grammatical analysis for the extraction of terminological noun phrases. In: Proc. Of 14th International Confer-ence on Computational Linguistics (COLING), pp. 977-981, Nantes (1992)
2. Heid, U..:Extracting terminologically relevant collocations from Ger-man technical texts. In: Proceedings of the TKE '99 International Congress on Terminology and Knowledge Engineering, pp. 241-255, Indeks, Innsbruck (1999)
3. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: Klavans & Resnik, 1996.
4. Chung, T.: A corpus comparison approach for terminology extraction. Terminology 9(2), pp. 221-246 (2003)
5. Drouin, P. and Doll, F.: Quantifying Termhood through Corpus Com-parison. In: Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds) : Managing Ontologies and Lexical Resources, pp. 191-206. Copenha-gen (2008)
6. Kit, C and Liu, X.: Measuring mono-word termhood by rank differ-ence via corpus comparison. Terminology 14(2), pp. 204-229 (2008)

7. Gémar, J.-C.: Traduction spécialisée et droit. Langages du droit, styles et sens. In: M. Gotti & S. Šarcevic, Insights into Specialized Transla-tion, pp. 79-106. Bern: Peter Lang (2006)
8. Kjær, Anne Lise: Phrasemes in legal texts. In: Harald Burger et al. (eds): Phraseology An International Handbook of Contemporary Research. (eds). Walter de Gruyter, pp. 506-516 (2007)
9. Nielsen, S.: The Bilingual LSP Dictionary: Principles and Practice for Legal Language. Narr Verlag (1994)
10. De Hertog, D., K. Heylen, and H. Kockaert.: A Variational Linguistics Approach to Term Extraction. In: Ú. Bhreathnach and F. de Barra Cusack (eds): Proceedings TKE 2010: Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges. Dublin, pp. 229-248 (2010)
11. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6(1), pp. 97-133 (2001)
12. Labov, W.: Some Principles of Linguistic Methodology. Language in Society, 1: pp. 97-120 (1972)
13. Speelman, D., Gondelaers, S. and Geeraerts, D. : A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. In: A.Wilson, 17. D. Archer and P. Rayson (eds.), Corpus Linguistics around the World, pp. 195-202. Amsterdam: Rodopi (2006)
14. Scott, M.: Pc analysis of key words - and key key words. System 25, pp. 233-245 (1997)
15. Pedersen, T.: Fishing for exactness. In Proceedings of the South-Central SAS Users Group Conference, Texas (1996)
16. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp. 61-74 (1993)