IOP Conference Series: Earth and Environmental Science

PAPER • **OPEN ACCESS**

A Corpus-based Analysis of the Terminology of the Social Sciences and Humanities

To cite this article: Susi Yuliawati *et al* 2018 *IOP Conf. Ser.: Earth Environ. Sci.* **175** 012109

A Corpus-based Analysis of the Terminology of the Social Sciences and Humanities

Susi Yuliawati[1], Totok Suhardijanto[2] and Rahayu Surtiati Hidayat[2]

[1]Universitas Padjadjaran, Kampus Jatinangor, Jawa Barat, 45363. Indonesia
[2]Universitas Indonesia, Kampus Depok UI. 16424. Indonesia
Telp. +62217863528/+62227796482
E-mail: susi.yuliawati@unpad.ac.id

Abstract: This paper was concerned with terminology of the social sciences and humanities in Indonesian scientific papers. Using electronic corpora built from the collection of texts on legal science and administrative science in Universitas Indonesia, the aims of the study was to illustrate how to integrate corpus linguistic method with the communicative theory of terminology (CTT) to examine terminology. Three procedures of corpus analysis were applied to assist in identifying two of the areas that terminological units must fulfil, i.e. the linguistic component and the cognitive component. The keyword and word clusters analysis were used to extract multi-word terms while the collocation analysis was used to derive the most significant sense categories of terms. Using corpus software, vis. WordSmith Tools, the corpus analysis presents some of the results showing that the linguistic component of terminological units can be traced through the technique of keyword and word clusters. In Addition, the cognitive component of terminological units was possible to investigate through the concept of semantic preference, one of the key concepts in corpus linguistics built from the analysis of collocation. Therefore, it can be concluded that a corpus-based approach to study terminology is considered to offer several benefits, especially to the activity concerned with the compilation, description, processing and presenting terms in a more reliable and efficient way. It may also provide an alternative method for creating glossary and for translators to resolve terminological problems.

Keywords: legal science, administrative science, a corpus-based study, linguistic component, cognitive component, communicative theory of terminology.

1. Introduction

This paper presents a set of procedure for studying Indonesian terminology of the social sciences and humanities, more specifically in the fields of legal science and administrative science.The terminology is extracted from the collections of scientific papers at Universitas Indonesia that may give the big picture about terms used in the fields of law and administrative sciencein Indonesia. The research aims at identifying the terminological units by using corpus linguistic approach. The terminological units are examined through several techniques in corpus analysis, such as keyword and collocation analysis. The present paper assumes that a general word might become a term in a specific domain. In other words, there is a possibility that non-specialists consider a word to be a term, which is, however, only a general word for the specialists. Equally, it is possible that specialists use terms that their non-specialist audience takes to be words in the general language.

Therefore, the division between the general words and terms is not firm. Furthermore, it is stated that there is no different between terminological units and other linguistic units such as words or lexical units in a general usage because principally general and specialisedlanguage can be accommodated within one natural language. The only difference between terms and other lexical units reside in the fact that they fulfilrestricted conditions in each of their cognitive, grammatical, and pragmatic constituent elements. Consequently, lexical units have a potential to be a term and non-term and to legitimize them as a specific objects of terminology, it is necessary to demonstrate that they are specific and explain their specificity based on the triple composition of terminological units, viz. a linguistic component, a cognitive component and a communicative component.

The present study is a part of previous research, which is projected to provide a comprehensive and systematic analysis about the collection and description of terms. However, it addresses particular research questions that examineterms extractionfrom the collection of academic textson legal scienceand administrative scienceby using corpus linguisticmethod. First, it aims to identify terminological unitsbased on the linguistic component through keyword analysis. Second, it aims to recognize the terminological units based on the cognitive component through semantic preference analysis. The study of terminological units based on the communicative component is not part of the discussion in this researchbecause the component tends to focus on the function of terminological units as discourse units to identify individual as members of professional group and allow them to interact, communicate and transfer their knowledge as well. It means that the communicative component is rather difficult to explore based on textual analysis.

2. Methodology

The research uses the methodology that may be referred to as a *corpus-based terminolog*y. It is defined as “a working method that explores a collection of domain-specific language materials (corpus) to investigate terminological issues”. The electronic text corpora are used here to identify terms and to provide evidence for the usage of terms.

The analysis of the corpus conducted in accordance with the purpose of the study is assisted by corpus software, viz. WordSmith Tools, which has three main modules, i.e. wordlist, keyword, and concord. In this research, wordlist is used to generate two specialized corpora, i.e. the corpus of legal science and the corpus of administrative science; and one reference corpus, i.e. the reference corpus of social sciences and humanities. The module of keyword, which provides a word that occurs in unusual frequency in a given text by comparison with a reference corpus of a same kind, is used to generate keywords of the corpus under investigation. The keywordsare lexical words that may indicate of the ‘aboutness’ of the texts and are a valuable basis for examining specialised corpora. For the present research, the keywords are generated from the comparison of each of the specialised corpus (the corpus of legal scienceand administrative science) and the reference corpus (the corpus of social science and humanities). The module of concord is used to generate concordance or also known as key word in context (KWIC). Concordance displays every instance of specified word or other search term in a corpus with a given of proceeding and following context for each result, so it allows users to look at words in context.

To study the terminology of legal science and administrative science, the research uses corpus analysis to examine three areas, i.e. linguistic, cognitive and communicative components that terminological units must necessarily cover. From linguistic point of view, terms are lexical units in special fields that can be expressed by nominal categories or other lexical categories (verbs, adjective, and phrases) or other types of units: supralexical (specialised phraseology or fixed sequence) or infralexical (specialised formants). A technique from corpus analysisused to examine the linguistic component is keyword cluster. The keyword cluster that represents two or more words found repeatedly together in each others’ company, in sequence is necessary to identify since terms are frequently found in compound words instead of single words.

From cognitive point of view, terms depend on a thematic context; they occupy a precise place in a conceptual structure; and their specific meaning is determined by their place in this structure. To identify terms from this point of view, semantic preference, one of the key concepts in corpus linguistic referring to the relation, not between individual words, but between a lemma or word-form and semantically related words is used to examine the sense categories of term candidates.

From the perspective of communicative component, terms occur in specialised discourse and they adapt to this type of discourse to their thematic and functional characteristics. Terms are also regarded as discourse units that identify individuals as members of a professional group and make them possible not only to communicate and interact, but also to transfer their knowledge. Consequently, the terms funcion to tranfer knowledge.

Principally, the linguistic, cognitive and communicative components are inseparable for a comprehensive description of terminology. However, the three approaches are separately examined in this research to show how corpus linguistics can be used to study terminology, more specifically from the communicative theory of terminology (CTT) proposed by Cabré.

3. Results and Discussions

The collection of texts from doctoral dissertations on legal scienceand administrative sciencein Universitas Indonesia, as the sample of Indonesian scientific papers on social sciences and humanities, is used to build two specialised corpora. The corpus of legal scienceis constructed from 17 doctoral dissertations and the corpus of administrative scienceis constructed from 7 doctoral dissertations. One reference corpus is built from the collection of 89 doctoral dissertations on economics and business, psychology, social and political sciences, humanities, legal science, and administrative science, which provide background data for reference comparison when generating keyword. Using WordSmith Tools, the word list of the legal science corpus is generated, which consists of 1.153.057 words while the administrative science corpus consists of 430.648 words. The word list of the reference corpus consists of 6.597.975 words.

The following analysis is meant to illustratehow the method of corpus linguisticscan be used to investigate two components that terminological units must fulfil to show their specificity. From the perspective of linguistic component, terminological units need to fulfil several conditions, for example, they are lexical units in special fields in the form of word or phrases. In relation to this, we argue that one of the techniques in corpus method that can be used to identify term candidates in special fields is keyword analysis. The keyword analysis that aims to find out which words characterize the text under investigation may be indicative of either what the text is about or what words are important.This procedure is regarded suitable to extract term candidates.The present paper is, however, particularly interested in examining terms in the form of multi-word units since termsare frequently compound words instead of single word.Therefore, the keyword list, which provides overview about the main subject in the text, is regarded as starting point for further analysis,especially in connection with the calculation of word clusters.

Word clusters, also known as lexical bundles, are sequences of words showing a statistical tendency to co-occur in a particular register. Even though word clusters create a tighter relationship than collocation, they simply represent repeated strings that may or may not prove to be the case of true multi word-units. Thus, list of word clusters from keywords generated by WordSmith Tools needs to be analysed further.

Table 1. The Keyword Clustersof the Corpus of Legal Science (terms candidates).

Keyword Cluster	Frequency
perguruan tinggi ‘higher education’	743
arbitrase ICSID ‘arbitrage ICSID’	372
LINGKUNGAN HIDUP *‘living environment’	313
perkawinan campuran ‘intermarriage’	304
penanaman modal ‘capital investment’	251
wajib pajak ‘taxpayer’	245
modal asing ‘foreign capital’	233

jabatan presiden ‘presidency’	224
tindak pidana ‘criminal act’	222
pengguguran kandungan ‘abortion’	219

To extract terms candidates, the procedure of analysis begins with creating a list of keyword cluster. Using WordSmith Tools, for example, the keyword clusters of the corpora of legal scienceand administrative science are created by comparing the word lists of each of the corpus with the reference corpus of social sciences and humanities. In the order of frequency of occurrence, Table 1 and 2 display the top ten keyword clusters of legal science and administrative science corpora. Most of them are found in the form of noun phrase (lexical units), which indicates that they fulfil the condition as the linguistic component.

Table 2. The Keyword Clusters from the Corpus of Administrative science (terms candidates)

Key word Cluster	**Frequency**
KDH tingkat‘head of region level’	796
bupati KDH ‘regent the head of regency’	664
walikotamadya KDH ‘mayor the head of city’	571
*BUDAYA PERUSAHAAN**‘corporate culture’	386
pemerintah pusat‘central government’	315
penyerahan wewenang‘transfer of powe’	268
RISTEK Industri‘research and technology of industry’	239
gaya manajemen‘management styles’	209
kinerja karyawan‘employee performance’	184
cara penyerahan‘method of submission’	182

The keyword clusters here, however, are still regarded as term candidates because they must be examined from two other components, i.e. the cognitive component and the communicative component. It is because the fact that some of words or phrases are possible to be a part of terminology in several fields of study. For example, the word clusters from the legal sciencecorpus, such as PERGURUAN TINGGI, *LINGKUNGAN HIDUP*, and PENANAMAN MODAL may also be terms in the fields of education, ecology, and economics respectively. Another example, word clusters from the administrative science corpus, such as *BUDAYA PERUSAHAAN* and GAYA MANAJEMEN, could also be terms in business and management sciences.

To determine whether term candidates derived from the keyword cluster analysis can be classified as terms in the fields of law and administrative science, further investigation from the perspective of cognitive component needs to be done. In this analysis, the keyword cluster *LINGKUNGAN HIDUP* from legal science corpus and *BUDAYA PERUSAHAAN* from the administrative science corpus will be analysed further to illustrate how the cognitive componentof terminological units can be examined using corpus method. From the cognitive point of view, terminological units are required to follow several restricted conditions. The particular conditionthat is possible to explore by using corpus linguistic perspective is that they depend on a thematic context, which indicates that their specific meaning is determined by their place in certain conceptual structure and context. Semantic preference, a concept built upon a collocation analysis [9],extracts meaning arising from the common semantic features of collocates of a given node item [14]. The concept allows us to examine the specific meaning of term candidates based on habitually co-occurring words, which share semantic features, using a statistical measure. In this analysis, two examples of term candidates taken from the list of keyword clusters are analysed by semantic preference to demonstrate their specific meaning in the related field of study. The first procedure to do this is determining significant collocates of the node items, i.e. *LINGKUNGAN HIDUP* 'living environment' and *BUDAYA PERUSAHAAN* 'corporate culture'.To derive significant collocates computationally, the parameters used are MI score of 3.00 or higher and the minimum frequency of 5 with a span of 4:4 around the node.

Table 3. The Semantic Preference of *LINGKUNGAN HIDUP* .

Semantic category	**Significant collocates**
crime, law and order	*hukum, peraturan, ketentuan-ketentuan, ketentuan, undang-undang, perundang-undangan,UULH, UU, sanksi, pasal, mengatur, pengaturan*
green issues	*pencemaran, pelestarian, pengelolaan, perusakan, kelestarian, konservasi, kerusakan, perlindungan, alam, sumber daya, lingkungan, menjaga, memelihara, menanggulangi, pembangunan, masalah, menyebabkan, kesadaran, fungsi, kemampuan, menunjang*
people	*manusia, masyarakat, orang, kependudukan*
government	*ASEAN, nasional, Indonesia, negara, menteri, daerah*

The analysis of the keyword cluster *LINGKUNGAN HIDUP* 's significant collocates reveals five categories of semantic preference, as shown in Table 3. It indicates that the most important sense categories for *LINGKUNGAN HIDUP* in the corpus of the science of law are crime, law and order; green issues; people; and government. There are some sense categories that can be seen as something specific. For example, *LINGKUNGAN HIDUP* that co-occurs with a set of semantically related words, such as *hukum* 'law', *peraturan* 'regulation', *ketentuan-ketentuan* 'provisions', *ketentuan* 'provision', *undang-undang* 'act', *perundang-undangan* legislation, *UULH* 'Environmental Act', *UU* Act'', *sanksi* 'punishment', *pasal* 'article', *mengatur* 'regulate', and *pengaturan*'regulation' demonstrates that the use of *LINGKUNGAN HIDUP* is strongly associated with legal system and criminal activities. The set of collocates building the semantic preference of crime, law and order for the keyword cluster *LINGKUNGAN HIDUP* is large and significant enough, which indicates that the keyword clusteris used in a specialised domain, i.e. law science instead of ecology or biology. Furthermore,the specialised domain of *LINGKUNGAN HIDUP* is affirmed by the semantic preference of green issues. It tends to discuss environment in relation to the violation of law (realizedin the collocates of *lindungan* 'protection' *konservasi* 'conservation', *menjaga* 'preserve', and *memelihara* 'maintain').

Another *pencemaran* 'pollution', *perusakan* 'destruction', *kerusakan* 'damage' and *masalah* 'problems') and the function of law (seen from the collocates of *pelestarian* 'preservation', *kelestarian* 'sustainability', *per*set of semantic preference showing that *LINGKUNGAN HIDUP* is not merely used to refer to surroundings or conditions in which a person, animal or plant lives or operate scan be seen from the sense category of government. The keyword cluster *LINGKUNGAN HIDUP*, which co-occurs with the words *ASEAN* 'Association of Southeast Asian Nations', *nasional* 'national', *Indonesia, negara* 'country', *menteri* 'minister', and *daerah* 'regional', is used to talk about the role of governments in various levels, regional, nationaland international, in regulating and protecting environment. Based on the analysis of semantic preference, it has proven that *LINGKUNGAN HIDUP* is used in a specific thematic context, i.e. legal science, because it is not only associated with green issues and people, but also with the system of rules and the role of government in environmental protection. Thus, *LINGKUNGAN HIDUP* can be regarded as a term in the science of law.

In the case of *BUDAYA PERUSAHAAN*, the analysis of its significant collocates reveals nine categories of semantic preference, as shown Table 4. It demonstrates that the most important sense categories for *BUDAYA PERUSAHAAN* in the corpus of administrative scienceare power, organizing; affect; people; group; mental action: thought, belief; mental object: conceptual object, means, method; planning; comparing; and importance. Some of the sense categories strongly indicate that the keyword cluster *BUDAYA PERUSAHAAN* has a specific meaning. First, it can be seen from the co-occurrence of *BUDAYA PERUSAHAAN* with the collocates *sikap* 'attitude' and *keyakinan* 'belief', which share the semantic feature of mental action, and with the collocates *konsep* 'concept', *cara* 'means/ways', and *pola* '*pattern*', which share the semantic feature of mental object. These sets of semantic preference apparently convey the concept of culture of the keyword cluster *BUDAYA PERUSAHAAN*, but then the concept becomes more specific which can be seen from the semantic preference of power, organizing. The co-occurrence of *BUDAYA PERUSAHAAN* with the collocates *gaya manajemen* 'management style', *perusahaan* 'corporate', *daya* 'power' and *manajemen* 'management' specifies that the concept of culture talked about is related to the process or activity of running organization or business.

Table 4. The Semantic Preference of *BUDAYA PERUSAHAAN*.

Semantic category	**Significant collocates**
power, organizing	*gaya manajemen, daya, perusahaan, manajemen*
affect	*hubungan, pengembangan, faktor, adaptasi, adaptif, mengembangkan, kausal, regression, faktor-faktor, penyebab*
people	*karyawan, warga, responden, manajer*
group	*tim,bersama*
mental action: thought, belief	*sikap, keyakinan*
mental object: conceptual object, means, method	*konsep, cara, pola*
planning	*strategi, tujuan*
comparing	*variasi, kecenderungan, variabel*
importance	*nilai, nilai-nilai, signifikan*

In addition, another set of semantic preference that constructs the specific meaning of *BUDAYA PERUSAHAAN* is identified from the co-occurrence of *BUDAYA PERUSAHAAN* with the collocates *hubungan* 'relation', *pengembangan* 'development', *mengembangkan* 'to develop', *regression, faktor* 'factor', *faktor-faktor* 'factors', *kausal* 'causal', *penyebab* 'cause', *adaptasi* 'adaptation', and *adaptif* 'adaptive', which share the semantic feature of affect. The semantic preference demonstrates that *BUDAYA PERUSAHAAN* is frequently discussed in connection with the cause and the change that they brings for organization or business. This analysis of semantic preference has revealed that the keyword cluster *BUDAYA PERUSAHAAN* is used in a specific context and thus has a specific meaning. The sets of semantic preference constitute the concept of *BUDAYA PERUSAHAAN* representing the specificity of knowledge in the field of administration as it closely associated with process of managing and developing organization or business. Consequently, the keyword cluster of BUDAYA PERUSAHAN can be regarded as one of the terms in the field of administrative science.

4. Conclusion

Based on the result of analysis, there are several essential points that can be concluded about the term extraction from the electronic corpora of social sciences and humanities. First, terms derived from the collection of large quantity of academic texts has proven that key word and word clusters analysis can be used an alternative method to extract multi-word terms. Second, semantic preference, built from the relation between a word and semantically related words, allow us to extract and highlight the most significant sense categories a term has. In other words, sets of semantic preference of a term constitute conceptual unit representing nodes of knowledge, which is relevant to the field of subject under investigation. In other words, the concept of semantic preference provides us a method to examine the actual use of a term in a variety of contexts. Hence, it assists us gathering information to derive and to comprehend more comprehensively the specific meaning of a term. A corpus-based approach to study terminology is considered to offer several benefits, especially over the traditional paper-based approach. Since it allows researchers to investigate terms and concepts from big quantity of data, it is regarded to contribute to the activity concerned with the compilation, description, processing and presenting terms in a more reliable and efficient way. It may also provide an alternative method for creating glossary and for translators to resolve terminological problems.

5. References

[1] Sager, J.C. (1990). *A Practical Course in Terminological Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

[2] Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

[3] Cabré, M.T. (2003). *Theories of Terminology: Their Description, Prescription and Explanation. Terminology 9:2*

[4] Cabré, M.T. (1999). *Terminology: Theory, Methods and Application*. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing Company.

[5] Kast-Aigner, J. (2009). *Terms in context: A corpus-based analysis of the terminology of the European Union's development cooperation polic*y. *Fachsprache, International Journal of Specialized Communication,* Vol. XXXI 3-4, pp. 139–152.

[6] Scott, M. (1997). *PC analysis of key words – and key key words. System,* Vol. 2 No. 2, pp. 233-245.

[7] Baker, P., Hardie, A. & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburg: Edinburg University Press.

[8] Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University.

[9] McEnery, T. & Hardie, A. (2012), Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

[10] Cabré, M.T. (2010). *Terminology and translation.* In Gambier, Y. &Doorslaer, L.V. (Eds.). *Handbook of translation studies volume 1.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

[11] Scott, M. (2013). *WordSmith tools manual.* Liverpool: Lexical Analysis Software, Ltd.

[12] Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. UK & USA: Blackwell Publishing.

[13] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, Essex: Pearson Education, Ltd.