# Collocations and Grammatical Patterns in a Multilingual Online Term Bank

Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, South Africa (elsabe.taljard@up.ac.za)*

**Abstract:** This article considers the importance of including various types of collocations in a terminological database, with the aim of making this information available to the user via the user interface. We refer specifically to the inclusion of empirical and phraseological collocations, and information on grammatical patterning. Attention is also given to provision of information on semantic prosody and semantic preferences — aspects which have been rather neglected in South African terminological databanks and language for special purposes (LSP) dictionaries. Various strategies for the effective semi-automatic extraction of collocational data from specialized corpora are explored. Possibilities regarding access to and presentation of collocational information to the user are briefly considered. It is argued that users should have direct access to collocational information, and that collocations should not only be accessible via the lemmatic address of the term appearing as part of the collocation. The research is done within the context of the establishment of an Open Access Resource Term Bank, which is developed as a pedagogical tool to support students whose language of learning and teaching is not the L1.

**Keywords:** COLLOCATIONS, GRAMMATICAL PATTERNING, MULTILINGUAL TERMINOLOGY DATABASE, SEMANTIC PROSODY, SEMANTIC PREFERENCE, OPEN ACCESS RESOURCE TERM BANK, CORPUS-BASED TERMINOLOGY

**Opsomming: Kollokasies en grammatikale patrone in 'n veeltalige aanlyntermbank.** Hierdie artikel ondersoek die belangrikheid van die insluiting van verskillende tipes kollokatiewe inligting in 'n veeltalige terminologiese databank met die oog daarop om hierdie inligting aan die gebruiker via die gebruikerskoppelvlak beskikbaar te stel. Ons verwys spesifiek na die insluiting van emipriese en fraseologiese kollokasies, en ook na inligting oor grammatiese patrone. Aandag word ook gegee aan die verskaffing van inligting oor semantiese prosodie en semantiese voorkeur — aspekte wat nie veel aandag in Suid-Afrikaanse terminologiese databanke en vakwoordeboeke geniet nie. Verskeie strategieë vir die effektiewe semi-outomatiese onttrekking van kollokatiewe data uit vakgerigte korpora word ondersoek. Verskillende moontlikhede met betrekking tot die toegang tot en aanbieding van kollokatiewe inligting aan die gebruiker word oorweeg. Daar word geargumenteer dat gebruikers direkte toegang tot kollokatiewe inligting moet hê, en nie alleenlik via die lemmatiese adres van die term wat as deel van die kollokasie optree nie. Die navorsing word gedoen teen die agtergrond van die daarstel van 'n terminologiese hulpbron met vrye toegang, wat as opvoedkundige hulpmiddel ontwerp word ter ondersteuning van studente wat nie deur middel van hul L1 studeer en/of onderrig word nie.

**Sleutelwoorde:** KOLLOKASIES, GRAMMATIESE PATRONE, VEELTALIGE TERMINOLOGIEDATABASIS, SEMANTIESE PROSODIE, SEMANTIESE VOORKEUR, VRY-TOEGANKLIKE TERMINOLOGIESE HULPBRON, KORPUS-GEBASEERDE TERMINOLOGIE

## Contextualization and introduction

The research reported on is done within the context of the establishment of a multilingual, open education resource term bank (OERTB). Establishment of such a term bank is part of a Department of Higher Education funded project, awarded jointly to the University of Pretoria and the University of Cape Town. It is envisaged as a collaborative effort between all South African universities and the aim is to create a terminological tool which can serve as pedagogical support tool to South African students. This tool will be made available to participating universities under a creative commons licence. Access to the user interface will be via the online learning systems of the various universities. The assumption is that the majority of South African students are exposed to a tertiary education system where the language of teaching and learning is not the strongest language, i.e. the L1. It is furthermore assumed that giving these students access to an internet-based term bank, which contains not only term equivalents for key concepts in the African languages, but also additional conceptual information, e. g. definitions, and guidance on usage of terms, should aid in the conceptualisation of subject content. The tool is planned as an organic one, with terminology being developed within actual pedagogical situations.

In terms of Bergenholtz and Bothma (2011: 61, 62) we envisage that our terminological tool will be used in cognitive and communicative situations. They describe cognitive situations as knowledge seeking situations which are unrelated to specific usage situations such as text reception. Within a cognitive situation the user simply wants to find knowledge, which can be stored for later use. Term banks are listed as one of the most commonly used tools in these situations. Communicative situations deal with problems or doubts that the user may have regarding the process of oral or written communication, and with issues such as text reception, text production, translation and text correction, of which the first three are possibly the most important in our specific usage situation. The practical implementation of our terminological tool deals directly with the communicative function in that it can provide a starting point for translanguaging, a practice which is described by Park (2013: 50) as assisting "multilingual speakers in making meaning, shaping experiences, and gaining deeper understandings and knowledge of the languages in use *and even of the content that is being taught*" (my emphasis). Within the context of tertiary education in South Africa forums such as tutorials would present ideal opportunities for translanguaging practices. In these pedagogical situations, students can discuss threshold concepts in their L1/strongest language, and draw on the terminological tool for African language equivalents of the English terms and explanations of key concepts in their L1.

The aim of this article is first to critically consider the importance of the inclusion of two types of information in a multilingual terminological database, with the aim of making this information available to the user via the user interface. We refer here to various kinds of collocational information, which contribute to the conceptual information provided in the term bank, and grammatical patterning, which is more usage-oriented. Pending a detailed discussion of the notion of collocation (see below), it can provisionally be described as frequently recurring word combinations. Secondly, various strategies are investigated for the effective extraction of collocational information and grammatical patterning from electronic text corpora, within the time and resource constraints of the project. Reference to time and resource constraints here may seem redundant, but within the specific context these constraints are indeed relevant. We are aware of the fact that very sophisticated procedures and tools for, inter alia, computational identification and extraction of collocations exist; however, the level of computational expertise required to utilize these resources is far beyond what is realistically available within the constraints of the project. The project is funded for three years and expenditure is limited to a fixed budget. The project team therefore has no choice but to make use of commercially available software, even though we are aware of their limitations. In the last instance, different possibilities regarding access to and presentation of collocational information to the user are briefly considered.

For the sake of clarity, a few remarks concerning the terminology used within the context of any kind of electronic terminology activity, tools or products are necessary. Perhaps somewhat ironically, the terminology used in this regard is rather confusing — the terms 'terminological/terminology data bank', 'term bank' and 'terminological database' being used rather indiscriminately and sometimes interchangeably, both generally and in scholarly work.

A trawl through the literature has brought the following to light: the terms 'term bank' and 'terminology/terminological data bank' are treated as synonyms and refer to a collection of different, but usually related databases that can be accessed by users with common software via a user interface (UI). A term bank usually belongs to or refers to an institution. Examples of well-known term banks are TERMIUM Plus® (https://www.btb.termiumplus.gc.ca) and *Grand dictionnaire terminologique* (GDT) (https://www.granddictionnaire.com) — two Canadian term banks — and InterActive Terminology for Europe (IATE). The OERTB would be an example of a term bank.

Definitions of the term 'database' emphasize the notion of a structured collection of terminological data, cf. the definition provided by the ISO Online Browsing Platform (OBF) (https://www.iso.org/obp/ui/, accessed 18-05-2015), according to which a terminology database consists of structured sets of terminological records in an information processing system. It is important to understand that users do not have access to the terminological database itself; they only have access to the information that the terminologist/databank manager chooses to make available via the user interface. A second important point is that the database may (and usually does) contain many more data categories than the ones which are accessible to the user. The current research investigates the inclusion of collocational information and information on grammatical patterning in the OERTB database with the aim of making this information available to the user through the user interface.

## On defining collocations and grammatical patterning

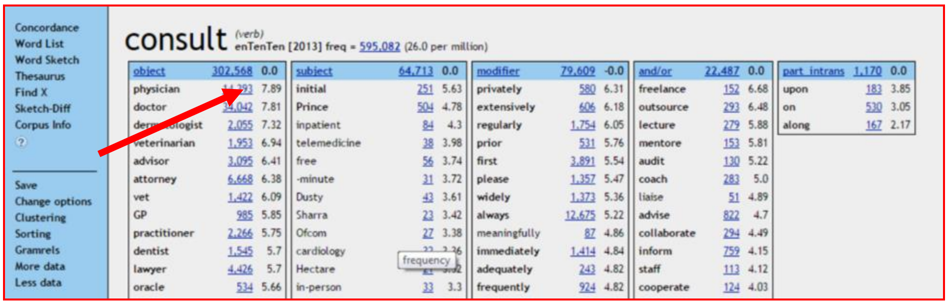
The concept of collocation is notoriously difficult to define, even though, as Evert (2007) points out, it is based on a widely-shared intuition that certain words have the tendency to co-occur in natural language. From the literature it is clear that collocation is a multi-dimensional phenomenon. Evert (2007) distinguishes between the Firthian and the phraseological interpretations of the notion of collocation. Within the Firthian tradition collocation is the recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages, cf. Smadja (1993). They are observable facts about language and thus present primary data. When working with raw, i.e. un-annotated corpora, the notion of collocation as recurrent word combinations implies a kind of mechanistic lexical co-occurrence where the presence of a node primes the presence of a collocate or collocates: *salt* and *pepper*, *cow* and *milk,* and *day* and *night* are prototypical examples of what Evert calls empirical collocations; for the verb *request* object nouns that can be expected to co-occur are *information*, *permission*, *assistance* and *help*, the collocational pairs being *request* and *information*, *request* and *permission*, etc. The phraseological interpretation on the other hand, describes collocations as being semi-compositional and lexically determined word combinations, such as *make an appearance* and *give a talk*. This kind of collocation is also known as multiword expressions and includes a whole range of subcategories, from completely opaque idioms to combinations which are subject to arbitrary lexical restrictions, e.g. *take medicine* rather than *drink medicine*.

Although interesting from a theoretical point of view, the distinction between empirical collocations and multiword expressions is not of primary importance for our project, the reason being twofold. First, the available software programmes i.e. *SketchEngine* (http://www.sketchengine.co.uk) and *WordSmith Tools* (http://www.lexically.net/wordsmith) which are used for the computational processing of terminological data, do not distinguish between these two kinds of collocations. Secondly, both kinds are relevant to the current project. It would seem that empirical collocations are relevant on the conceptual level, in that members of a collocational set could be conceptually related, whereas phraseological collocations seem to be relevant more on a usage-related level. This issue will be further investigated below.

From a semantic point of view, collocation is represented by two related phenomena, i.e. semantic preference and semantic prosody, which both describe the statistically significant co-occurrence of a word with a group of other words, cf. Kübler and Pecman (2012: 188). Semantic prosody refers to the measure with which the (affective) meaning of a word is coloured by its typical collocates, whereas semantic preference is described as the measure of co-occurrence between a word and a set of semantically related words. To briefly illustrate the notion of semantic prosody: the fact that the typical (abstract) objects that collocate with the verb *tolerate* in the *enTenTen* corpus are *dissent*, *disrespect*, *nonsense*, *intolerance*, *mediocrity* and *harassment*, the overall negative implication of the collocates taints the meaning of *tolerate* as being negative, resulting in a negative prosody for *tolerate*. Since it is assumed that the meaning of a term has been previously delimited by means of a definition, and that terms are supposed to be emotionally neutral, the perception that has hitherto prevailed that semantic prosody would play a lesser role within terminological work is understandable.

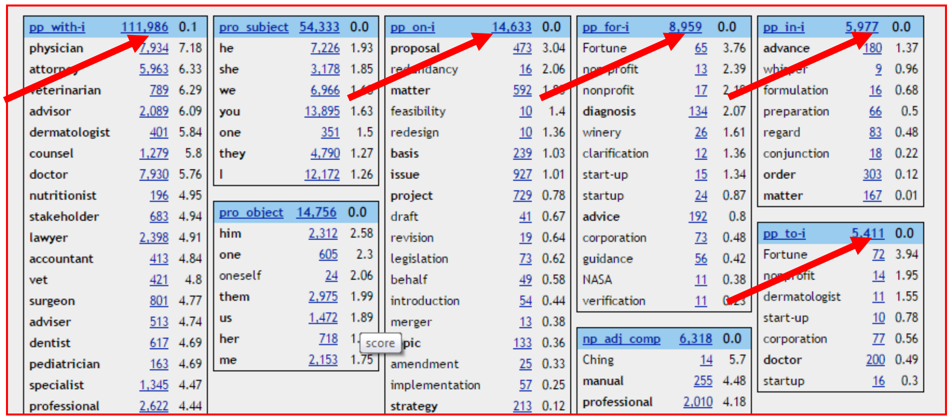
The following example — where the collocates of the verb *consult* are called up in a WordSketch, using the *enTenTen* corpus — is a good illustration of semantic preference. The top collocates for objects appearing with *consult* are *physician*, *doctor*, *dermatologist*, *veterinarian*, *advisor* and *attorney*, revealing the semantic preference of the verb *consult* to appear with objects sharing the semantic feature 'professional individual'. Kübler and Pecman (op. cit.) point out that semantic preference has recently aroused more interest in specialized languages, i.e. language for special purposes and has resulted in collocations being more commonly taken up in LSP dictionaries and term bases. They furthermore state that 'phenomena such as semantic prosody and preference would provide the user with complete and necessary information'.

Grammatical patterning represents another kind of collocation, i.e. the grammatical company that a word keeps. This is typically the kind of information which one would find in the WordSketches in SketchEngine. A WordSketch of the verb *consult* in the English *enTenTen* corpus, for example reveals that the grammatical pattern in which it most frequently appears is that of transitive verb, in which case it is followed by an object: the overall frequency of 'consult' as a transitive verb makes up almost 50% of its total occurrences. It also shows that the second most frequent grammatical pattern (24%) for 'consult' is to be followed by a prepositional phrase in which the preposition is most frequently 'with', followed by 'on', 'for', 'in' and 'to'. Compare the following screen shots from SketchEngine:



consult (verb) enTenTen [2013] freq = 595,082 (26.0 per million)

| object | 302,568 | 0.0 | subject | 64,713 | 0.0 | modifier | 79,609 | -0.0 | and/or | 22,487 | 0.0 | part_intrans | 1,170 | 0.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| physician | 14,393 | 7.89 | initial | 251 | 5.63 | privately | 580 | 6.31 | freelance | 152 | 6.68 | upon | 183 | 3.85 |
| doctor | 34,042 | 7.81 | Prince | 504 | 4.78 | extensively | 606 | 6.18 | outsource | 293 | 6.48 | on | 530 | 3.05 |
| dermatologist | 2,055 | 7.32 | inpatient | 84 | 4.3 | regularly | 1,754 | 6.05 | lecture | 279 | 5.88 | along | 167 | 2.17 |
| veterinarian | 1,953 | 6.94 | telemedicine | 38 | 3.98 | prior | 531 | 5.76 | mentore | 153 | 5.81 | | | |
| advisor | 3,095 | 6.41 | free | 56 | 3.74 | first | 3,891 | 5.54 | audit | 130 | 5.22 | | | |
| attorney | 6,668 | 6.38 | -minute | 31 | 3.72 | please | 1,357 | 5.47 | coach | 283 | 5.0 | | | |
| vet | 1,422 | 6.09 | Dusty | 43 | 3.61 | widely | 1,373 | 5.36 | liaise | 51 | 4.89 | | | |
| GP | 985 | 5.85 | Sharra | 23 | 3.42 | always | 12,675 | 5.22 | advise | 822 | 4.7 | | | |
| practitioner | 2,266 | 5.75 | Ofcom | 27 | 3.38 | meaningfully | 87 | 4.86 | collaborate | 294 | 4.49 | | | |
| dentist | 1,545 | 5.7 | cardiology | 22 | 3.36 | immediately | 1,414 | 4.84 | inform | 759 | 4.15 | | | |
| lawyer | 4,426 | 5.7 | Hectare | 21 | 3.32 | adequately | 243 | 4.82 | staff | 113 | 4.12 | | | |
| oracle | 534 | 5.66 | in-person | 33 | 3.3 | frequently | 924 | 4.82 | cooperate | 124 | 4.03 | | | |



| pp_with-i | 111,986 | 0.1 |
|---|---|---|
| physician | 7,934 | 7.18 |
| attorney | 5,963 | 6.33 |
| veterinarian | 789 | 6.29 |
| advisor | 2,089 | 6.09 |
| dermatologist | 401 | 5.84 |
| counsel | 1,279 | 5.8 |
| doctor | 7,930 | 5.76 |
| nutritionist | 196 | 4.95 |
| stakeholder | 683 | 4.94 |
| lawyer | 2,398 | 4.91 |
| accountant | 413 | 4.84 |
| vet | 421 | 4.8 |
| surgeon | 801 | 4.77 |
| adviser | 513 | 4.74 |
| dentist | 617 | 4.69 |
| pediatrician | 163 | 4.69 |
| specialist | 1,345 | 4.47 |
| professional | 2,622 | 4.44 |

| pro_subject | 54,333 | 0.0 |
|---|---|---|
| he | 7,226 | 1.93 |
| she | 3,178 | 1.85 |
| we | 6,966 | [illegible] |
| you | 13,895 | 1.63 |
| one | 351 | 1.5 |
| they | 4,790 | 1.27 |
| I | 12,172 | 1.26 |

| pro_object | 14,756 | 0.0 |
|---|---|---|
| him | 2,312 | 2.58 |
| one | 605 | 2.3 |
| oneself | 24 | 2.06 |
| them | 2,975 | 1.99 |
| us | 1,472 | 1.89 |
| her | 718 | 1.[illegible] |
| me | 2,153 | 1.75 |

| pp_on-i | 14,633 | 0.0 |
|---|---|---|
| proposal | 473 | 3.04 |
| redundancy | 16 | 2.06 |
| matter | 592 | [illegible] |
| feasibility | 10 | 1.4 |
| redesign | 10 | 1.36 |
| basis | 239 | 1.03 |
| issue | 927 | 1.01 |
| project | 729 | 0.78 |
| draft | 41 | 0.67 |
| revision | 19 | 0.64 |
| legislation | 73 | 0.62 |
| behalf | 49 | 0.58 |
| introduction | 54 | 0.44 |
| merger | 13 | 0.38 |
| topic | 133 | 0.36 |
| amendment | 25 | 0.33 |
| implementation | 57 | 0.25 |
| strategy | 213 | 0.12 |

| pp_for-i | 8,959 | 0.0 |
|---|---|---|
| Fortune | 65 | 3.76 |
| non-profit | 13 | 2.39 |
| nonprofit | 17 | [illegible] |
| diagnosis | 134 | 2.07 |
| winery | 26 | 1.61 |
| clarification | 12 | 1.36 |
| start-up | 15 | 1.34 |
| startup | 24 | 0.87 |
| advice | 192 | 0.8 |
| corporation | 73 | 0.48 |
| guidance | 56 | 0.42 |
| NASA | 11 | 0.38 |
| verification | 11 | 0.23 |

| np_adj_comp | 6,318 | 0.0 |
|---|---|---|
| Ching | 14 | 5.7 |
| manual | 255 | 4.48 |
| professional | 2,010 | 4.18 |

| pp_in-i | 5,977 | 0.0 |
|---|---|---|
| advance | 180 | 1.37 |
| whisper | 9 | 0.96 |
| formulation | 16 | 0.68 |
| preparation | 66 | 0.5 |
| regard | 83 | 0.48 |
| conjunction | 18 | 0.22 |
| order | 303 | 0.12 |
| matter | 167 | 0.01 |

| pp_to-i | 5,411 | 0.0 |
|---|---|---|
| Fortune | 72 | 3.94 |
| non-profit | 14 | 1.95 |
| dermatologist | 11 | 1.55 |
| start-up | 10 | 0.78 |
| corporation | 77 | 0.56 |
| doctor | 200 | 0.49 |
| startup | 16 | 0.3 |

**Figure 1:** WordSketch of *consult*

## Collocations and grammatical patterns in LSP dictionaries and terminological databases

In traditional paper LSP dictionaries collocations and information on the grammatical company that terms prefer are generally rather neglected, although as pointed out by L'Homme and Leroyer (2009), the addition of collocational information is regarded as extremely useful for specialized reference works, such as LSP dictionaries. Space constraints, coupled with the traditional view on terminology, i.e. that terms are context independent, are probably major contributing factors to the absence of these two information types in terminological products. With the advent of electronic lexicography, space is no longer of primary concern, but more importantly, the transition to corpus-based terminology not only provides access to huge amounts of data, but also opens up the possibility of semi-automatically extracting terminologically relevant data from corpora by means of corpus-query tools. The so-called modern approach to terminology furthermore places more emphasis on usage, with the use of real texts as primary sources of data, thus drawing on the importance of contextual information — which per definition includes collocational information — in satisfying the information needs of the user. The increased attention to collocational information in different kinds of terminology tools, whether it be online databases or electronic LSP dictionaries, can also possibly be ascribed to the influence of the work of Sinclair (2004) on extended units of meaning, and that of Hanks (2006), who argues that words only have meanings when they are put into context, thus establishing an association between word meaning and word use. A lexicographic application of Hanks' theory on Norms and Exploitation is the *Pattern Dictionary of English Verbs* (www.pdev.org.uk) in which users are offered prototypical syntagmatic patterns of meaning and use of each verb covered. This trend has also started to spill over into the design of a variety of terminology tools. The *E-Advanced Learner's Dictionary of Verbs in Science* (DicSci) project, reported on by Alonso et al. (2011), the ARTES project of which one outcome is the compilation of an online bilingual LSP dictionary, see Kübler and Pecman (2012), and the *Dictionnaire fundamental de l'informatique et de l'Internet* (DiCoInfo), described in L'Homme et al. (2012), are all examples of terminological tools in which collocational information is provided. Lastly, Fuertes-Olivera also pays extensive attention to collocations in his set of online Spanish–English *Accounting Dictionaries.*

## The importance of collocations in terminology in general, and for the OERTB project specifically

First, as pointed out by L'Homme (2006: 186), collocations are often unpredictable combinations, even in specialized language, and should therefore be treated in LSP dictionaries and/or term banks. This becomes especially important in a bilingual or multilingual situation where translation is one of the envisaged functions of the terminological tool. Collocations can potentially pose problems in translation, since they are often language specific and idiosyncratic. In Afrikaans for example, one 'picks up weight' (*tel gewig op*), whereas in English the verb which collocates with *weight*, is *gain*.

Secondly, collocations are domain dependent, which furthermore implies that collocations in general language with which the user may be familiar, may not apply in a specific subject field. For illustrative purposes, two small internet-based LSP corpora were compiled, one on climate change and one on film and drama studies. These were then queried by using SketchEngine (https:// the.sketchengine.co.uk). In both these corpora, the term 'atmosphere' was thrown up as a keyword and is thus regarded as a key concept in both these subject fields. A list of collocate candidates was then drawn up for 'atmosphere' in a language for general purpose (LGP) corpus (*enTenTen* corpus, 40 mil sample), and in the two LSP corpora respectively. The emphasis here is on finding conceptually related collocates, therefore grammatical formatives or function words which can hardly be said to represent subject specific concepts were disregarded, even though they make out a sizable portion of the top collocate candidates. Therefore, only collocates with lexical and therefore conceptual content are listed in the table. The collocational span was set at 5 positions to left and right of the search node; the statistical measure used is T-score, one of the options offered by the SketchEngine.

| **LGP corpus (enTenTen)** | **LSP corpus$_1$ (Climate change)** | **LSP corpus$_2$ (Drama and Film)** |
|---|---|---|
| *create* | *carbon* | *creates* |
| friendly | CO | *create* |
| relaxed | dioxide | upper |
| great | *Earth* | play (n) |
| place | greenhouse | render |

| | | |
|---|---|---|
| *Earth* | gases | gothic |
| *carbon* | increase (v) | studying |
| warm | methane | filming |
| *creating* | oceans | setting (n) |
| people | released | research (n) |

**Table 1:** Top ten raw collocates for 'atmosphere' in three different corpora

From the above, it is clear that the collocations for 'atmosphere' are indeed domain specific — there is no overlap between the collocates for 'atmosphere' in the two LSP corpora, and only a small overlap between the collocates found in the LGP corpus on the one hand, and those found in the two LSP corpora respectively. This is especially clear when looking for example at the verbal collocates of 'atmosphere' in the two LSP corpora. Verbs collocating with 'atmosphere' in the Climate change corpus are 'increase' and 'released', whereas 'create(s)', 'render', 'studying' and 'filming' are typical collocates in the Drama and Film corpus.

A third reason why collocations are important in an LSP environment is the fact that empirical collocations are assumed to be useful elements for conceptualizing a knowledge domain, as Fuertes-Olivera (2012) points out. Since the OERTB is especially aimed at assisting with conceptualization of key terms in different subject fields, this is an issue which needs special attention. Apart from providing straightforward collocational information, serious consideration should be given to provide information on the conceptual relationships existing between collocates by means of collocational networks, as described by Alonso, Millon and Williams (2011) and Williams (1998). Collocational networks are described as "statistically based chains of collocations, a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation" (Alonso et al. 2011: 15). Williams (1998) argues that concepts central to a specific subject field are related, and that similar relational patterns can be identified in their surface constructs, i.e. words, or in our case, terms. Therefore, the frame of reference for any term is to be found in the lexical environment within which it appears, and which is revealed through collocation. It is assumed that concepts can be better grasped when they are presented within the environment in which they are used.

Collocational networks can be revealed by further processing of collocational information. As a starting point the strength of the association between each of the top x number (according to the keyness score) of single and multiword terms and their collocates are calculated, using e.g. Mutual Information (MI) as statistical measure. MI determines the strength of the association between two words: in a given finite corpus MI is calculated on the basis of the number of times the pair is observed together versus the number of times they appear separately. Each of the top ranking terms forms the node of a collocational network, and each collocate in turn is regarded as a node of a new collocational network. The collocational network is thus extended up to the point where no more significant collocates are found. The end result would then be a network of related concepts, positioning each concept within the conceptual framework of at least a particular section of the special subject field.

The issue of empirical collocates is complicated by the fact that different statistical measures result in vastly different results. Compare the following table in which collocates for 'atmosphere' were drawn from the LSP corpus on climate change using MI (Mutual Information) score for the left-hand column and T-score for the right-hand column. Briefly explained, the difference between these two measures is as follows: The t-score is a measure not of the strength of association but the confidence with which we can assert that there is an association. MI is more likely to give high scores to totally fixed phrases whereas t-score will yield significant collocates that occur relatively frequently (http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html). The right-hand column is the raw collocate list, i.e. one in which the function words were retained:

| | Freq | MI |
|---|---|---|
| **travels** | 3 | 8.761 |
| **thicker** | 3 | 8.498 |
| **traps** | 3 | 8.275 |
| **amplify** | 4 | 8.176 |
| **heat-trapping** | 7 | 8.057 |
| **constantly** | 3 | 7.913 |
| **Winds** | 3 | 7.761 |
| **inert** | 3 | 7.761 |
| **coupled** | 4 | 7.591 |
| **overlying** | 3 | 7.498 |
| **composition** | 6 | 7.439 |

| | Freq | T-score |
|---|---|---|
| **the** | 428 | 19.512 |
| **in** | 214 | 14.02 |
| **.** | 191 | 12.447 |
| **of** | 168 | 11.855 |
| **,** | 164 | 11.108 |
| **and** | 106 | 9.178 |
| **into** | 76 | 8.648 |
| **to** | 90 | 8.48 |
| **2** | 66 | 7.917 |
| **carbon** | 64 | 7.876 |
| **is** | 73 | 7.85 |

**Table 2:** Collocates for 'atmosphere' in the Climate change LSP corpus according to MI and T-scores

According to the literature, cf. Clear (n.d.) and http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html, the choice of the measuring instrument for measuring the strength of collocational relationships depends also on the frequency of the items concerned. Clarifying the merits and the suitability of measuring instruments does not fall within the scope of this article, but is nevertheless something that needs to be clarified with experts. From the literature however, it would seem that MI is generally preferred in situations similar to the current one. Furthermore, it needs to be kept in mind that any list of collocates is only as good as the corpus it is based on, and it is possible that the two

LSP corpora compiled for the purposes of this article, namely to illustrate the value of collocational networks in LSP information sources are simply too small to render statistically significant results. (It needs to be mentioned here that within the parameters of the project we are planning on having at least 1 million word LSP corpora for each subject field.)

If, despite the concerns raised above, it is assumed that collocates which are generated by whatever statistical measure are indeed conceptually related to the search word, it is clear that sophisticated statistical processing such as reported on by Williams et al. (2012) would be necessary, and possibly also human intervention in the form of expert confirmation of conceptual relationships, to eventually present users with collocates in a format that satisfies their information needs. The typical format in which such a collocational network would be presented to the user, is a visual network, similar to what is found in Visuwords, a visual dictionary (http://visuwords.com/). The advantage of using such a visual network is that collocates can be visually represented, not in isolation, but as a complex network of semantic relationships which ultimately reveals their meaning, and thus aids with cognition — which is one of the main aims of the term bank. Currently, the possibility of adapting the software (TlTerm) which is used for the project to enable it to also generate such networks is being investigated. For the purposes of illustration, the term 'permafrost' and its collocational network as generated from the LSP corpus on climate change is used as an example:

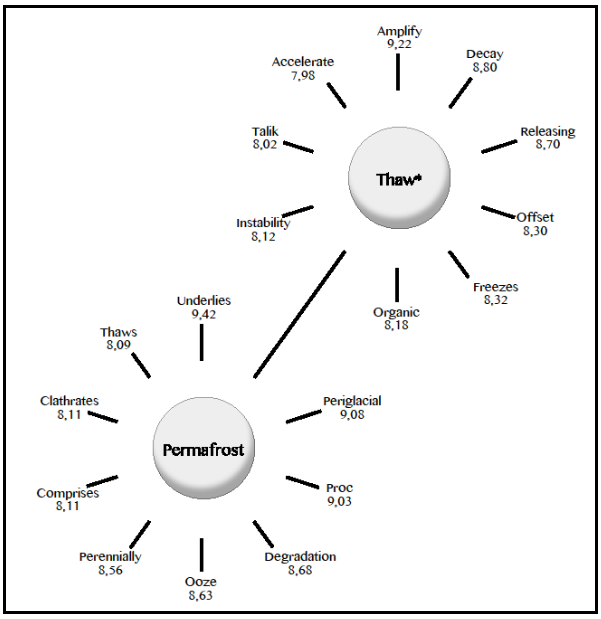
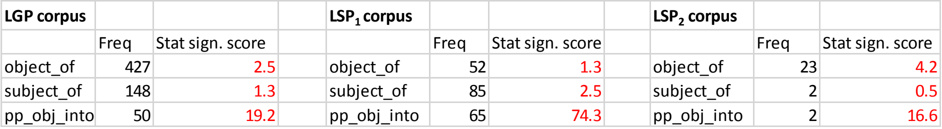


**Figure 2:** Collocational network for 'permafrost'

In the diagram above, the length of the lines connecting the different nodes are a reflection of the MI score, thus indicating the strength of association between the nodes. MI scores are displayed for illustrative purposes in the figure.

Further value can be added by specifying the type of relationship, e.g. 'kind of', 'part of', 'opposes' and 'is similar to' between collocates and a node and between different collocates. So, for example can the semantic relationship between 'thaw' and 'freeze' be indicated as being an antonymous one, etc. However, within the capacity, skills and time constraints of the OERTB project, this would be a rather ambitious undertaking. Even so, even if such an endeavour is not currently feasible, provision should be made for it in the conceptualization of the project, with a view to possible future implementation.

The last type of collocation to be discussed is grammatical patterning. Once again, this is important because (a) grammatical patterning in LGP is different from that of LSP, and (b) grammatical patterning also seems to be domain specific. Compare the following excerpt from a WordSketch generated for 'atmosphere' where the preference for specific grammatical patterns are revealed:



| **LGP corpus** | | | | **LSP$_1$ corpus** | | | | **LSP$_2$ corpus** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Stat sign. score | | | Freq | Stat sign. score | | | Freq | Stat sign. score |
| object_of | 427 | 2.5 | | object_of | 52 | 1.3 | | object_of | 23 | 4.2 |
| subject_of | 148 | 1.3 | | subject_of | 85 | 2.5 | | subject_of | 2 | 0.5 |
| pp_obj_into | 50 | 19.2 | | pp_obj_into | 65 | 74.3 | | pp_obj_into | 2 | 16.6 |

**Table 3:** WordSketch for 'atmosphere' across the three corpora

From the above it is clear that there is a bigger preference for 'atmosphere' to appear as the object of a verb in the Drama and Film corpus (LSP$_2$ corpus) than in any of the other two corpora. On the other hand, there does not seem to be a big tendency for it to appear as the subject of verbs in this corpus. Perhaps most significant is the preference for 'atmosphere' to appear as a prepositional object after 'into' in the Climate change corpus (LSP$_1$ corpus) — obviously because things are released or emitted into the atmosphere, or they escape into the atmosphere.

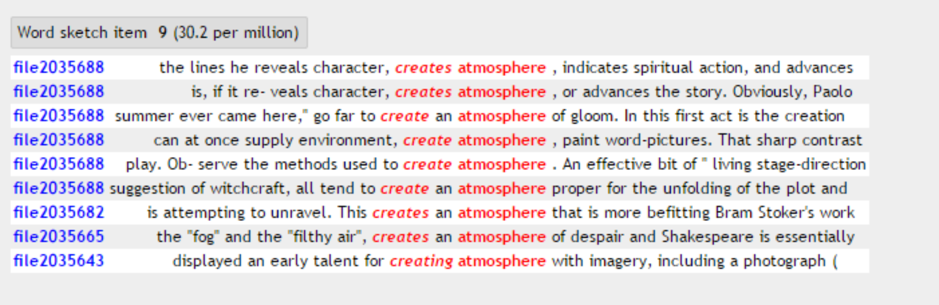
Apart from reflecting preference for grammatical patterns WordSketches also provide insight into which specific lexical items appear in these grammatical patterns, introducing a further level of collocational information. This excerpt reflects on a lexical level which modifiers tend to co-occur with 'atmosphere' in the various corpora:

It is clear that a relatively restricted set of modifiers co-occur with 'atmosphere' in the climate change corpus, with double the number in the drama and film corpus. Some of the modifiers appearing in the latter corpus also reveals something about the semantic prosody — the modifiers 'Gothic', 'mystical' (?), 'agoraphobic', 'chilling', 'deathly', 'cold', 'dark' reveal a clustering of modifiers which lend a negative prosody associated with atmosphere — information that may be useful to users, not necessary on a conceptual level, but rather on the pragmatic side.

| LGP corpus | | | | LSP$_1$ corpus | | | | LSP$_2$ corpus | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Stat sign. score | | | Freq | Stat sign. score | | | Freq | Stat sign. score |
| modifier | 676 | 1.5 | | modifier | 47 | 0.3 | | modifier | 19 | 1.2 |
| ---------------------------------------- | | | | ---------------------------------------- | | | | ---------------------------------------- | | |
| cozy | 10 | 8.27 | | Martian | 6 | 11.03 | | upper | 3 | 11.63 |
| relaxing | 8 | 8.08 | | upper | 7 | 10.95 | | Gothic | 2 | 10.75 |
| Martian | 5 | 7.81 | | inert | 2 | 10 | | genteel | 1 | 10.68 |
| homely | 5 | 7.78 | | standard | 2 | 9.66 | | mystical | 1 | 10.68 |
| friendly | 25 | 7.76 | | low | 7 | 8.78 | | agoraphobic | 1 | 10.6 |
| festive | 6 | 7.76 | | warm | 4 | 8.07 | | liberal | 1 | 10.54 |
| laid-back | 4 | 7.45 | | global | 2 | 6 | | chilling | 1 | 10.47 |
| controlled | 5 | 7.39 | | | | | | deathly | 1 | 10.47 |
| casual | 7 | 7.23 | | | | | | cold | 1 | 10.14 |
| intimate | 6 | 7.22 | | | | | | intense | 1 | 10.09 |
| oxygen-deficient | 3 | 7.18 | | | | | | northern | 1 | 9.95 |
| IDLH | 3 | 7.17 | | | | | | British | 1 | 9.71 |
| inert | 3 | 7.05 | | | | | | necessary | 1 | 9.02 |
| vibrant | 6 | 7.05 | | | | | | general | 1 | 9.02 |
| cosy | 3 | 6.99 | | | | | | dark | 1 | 8.89 |
| calming | 3 | 6.9 | | | | | | | | |
| ↓↓↓ | | | | | | | | | | |

**Table 4:** Modifiers co-occurring with 'atmosphere'

In the last instance, WordSketches provide yet another relatively simple procedure to retrieve collocations. In a WordSketch, items with a high frequency of occurrence within a particular grammatical pattern are clickable, thus revealing the concordance lines in which the search word, in this case 'atmosphere' appears together with its collocate. This could potentially provide additional information with regard to usage. Compare the following example of concordance lines containing 'create', which is one of the verbs which prefers 'atmosphere' as an object in the LSP$_2$ corpus:



Word sketch item 9 (30.2 per million)

file2035688 the lines he reveals character, creates atmosphere , indicates spiritual action, and advances
file2035688 is, if it re- veals character, creates atmosphere , or advances the story. Obviously, Paolo
file2035688 summer ever came here," go far to create an atmosphere of gloom. In this first act is the creation
file2035688 can at once supply environment, create atmosphere , paint word-pictures. That sharp contrast
file2035688 play. Ob- serve the methods used to create atmosphere . An effective bit of " living stage-direction
file2035688 suggestion of witchcraft, all tend to create an atmosphere proper for the unfolding of the plot and
file2035682 is attempting to unravel. This creates an atmosphere that is more befitting Bram Stoker's work
file2035665 the "fog" and the "filthy air", creates an atmosphere of despair and Shakespeare is essentially
file2035643 displayed an early talent for creating atmosphere with imagery, including a photograph (

**Figure 3:** Collocates for 'create' and 'atmosphere' in the LSP$_2$ corpus

What is noticeable in this example, is that when 'create' is used as the verb selecting 'atmosphere' as an object, the object is never preceded by a definite article. Apart from providing collocational information, these concordance lines therefore also give guidance with regard to usage. This specific example is particularly relevant for users who have an African language as home language. The use of articles in English is often problematic for such users, since the African languages do not distinguish the grammatical category 'article', consequently the use or non-use of articles in English often causes confusion due to language interference. Having access to this kind of information would assist users specifically with text production and translation.

## Presentation of and access to collocational information

With regard to the presentation of collocational information, two options are identified, i.e. implicit or explicit presentation. Implicit presentation would imply that the collocational data extracted from the LSP corpora according to the strategies described above would only be of value to the terminologist populating the database. In the case of explicit presentation, collocational information would be presented as such to the user, who will have the option of accessing this information by means of a search option. When collocational information is implicitly utilized, the terminologist would typically use these data to select usage examples which incorporate as much of the collocational data retrieved from the LSP corpus as possible. To illustrate: when selecting a usage example for the term 'atmosphere' within the drama and film domain, the terminologist will need to take the following into consideration:

— It is most often used as the object in a sentence (grammatical patterning)
— It frequently appears without an article, or with an indefinite one (grammatical patterning)
— The verb 'create' is one of the verbs which collocate with 'atmosphere'
— Many of the modifiers which co-occur with 'atmosphere' contribute to the expression of a negative semantic prosody.

These data would assist the terminologist in selecting an appropriate example from the corpus, e.g.

> The homes are cast with an unfriendly sterility that can create a chilling, agoraphobic ***atmosphere***

The disadvantage of treating collocations implicitly is that is does not provide sufficient collocational guidance — the user does not know whether *create a(n) atmosphere* is a frequent combination, or whether it is a mere coincidence. Users are oblivious to the fact that the example sentence illustrates both common usage of the term 'atmosphere' and a frequent collocation. Furthermore, no guidance is provided as to the negative semantic prosody which is often associated with the term 'atmosphere' in a drama and film context.

When presenting collocational information in an explicit manner, due consideration should be given as to whether this information should be displayed by default, in other words, whether on carrying out an initial search collocational information will automatically be displayed to the users, or whether they will have the option of accessing the information by means of a further search. Taking the target user of the OERTB as well as the function of the term bank into consideration, the designer of the user interface would be well-advised to heed the possibility of data overload. Collocational information should therefore be made available as an optional, additional search function which can be accessed by means of clicking on a dedicated button or tab. Care should furthermore be taken that the name of the button or tab which gives access to the collocational information should be transparent, making it clear to the users what kind of information they would find by clicking on it. Labelling such a button or tab as 'Collocations' would probably have little meaning for our envisaged users. Choosing a transparent label, such as 'Term in context' or 'Frequent combinations' rather than for example the neutral 'See more' has the advantage of providing the user with additional guidance to finding the required information.

Consideration should furthermore be given to allow users to directly access collocations, i.e. they must be able to search for a collocation, without having to access the collocation via the lemmatic address of the term appearing as part of the collocation. In the data base, collocations would then in effect be treated as multi-word terms, and can therefore be given the full treatment also given to single word lemmata.

## Conclusion

Provision of collocational data in any kind of terminological environment, whether it be LSP dictionaries or term banks has been sadly neglected, especially within the South African context. In this article it has been illustrated that in order to fulfil the information needs of the envisaged user, due consideration must be given to the provision of two kinds of collocational information, i.e. semantic prosody and semantic preference on the one hand, and grammatical collocation on the other. It has been illustrated that collocations are often unpredictable combinations, and since we envisage our term bank to also provide for translation needs, collocational information would be necessary. Collocations are furthermore domain specific and can assist with conceptualization within a particular subject field, thus forming a necessary component of information to be presented to the user. In order to reach this goal, full utilization must be made of data that can be extracted semi-automatically from corpora, and presented in such a way that it maximally satisfies the needs of the user. Easy access to collocational information is therefore of primary importance.

## Bibliography

*A Guide to Statistics.* n.d. Retrieved from http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html.

**Alonso, A, C. Millon and G. Williams.** 2011. Collocational Networks and their Application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci). Kosem, I. and K. Kosem (Eds.). 2011. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLEX 2011, November 2011, Bled, Slovenia*: 12-22. Ljubljana: Trojina, Institute for Applied Slovene Studies.

**Bergenholtz, H. and T.J.D. Bothma.** 2011. Needs-adapted Data Presentation in e-Information Tools. *Lexikos* 21: 53-77.

**Clear, J.** n.d. *t-score and Mutual Information Score from Birmingham Corpus Website*. http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html.

**Evert, S.** 2007. *Corpora and Collocations*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.6220&rep=rep1&type=pdf.

**Fuertes-Olivera, P.** 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28(1): 49-70.

**Hanks, P.** 2006. Metaphoricity is Gradable. Stefanowitsch, A. and S.Th. Gries (Eds.). 2006. *Corpus-based Approaches to Metaphor and Metonymy*: 17-35. Berlin: Mouton de Gruyter.

**Kübler, N. and M. Pecman.** 2012. The ARTES Bilingual LSP Dictionary: From Collocations to Higher Order Phraseology. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 187-210. Oxford: Oxford University Press.

**L'Homme, M.-C.** 2006. The Processing of Terms in Dictionaries: New Models and Techniques. A State of the Art. *Terminology* 12(2): 181-188.

**L'Homme, M.-C. and P. Leroyer.** 2009. Combining the Semantics of Collocations with Situation-driven Search Paths in Specialized Dictionaries. *Terminology* 15(2): 258-283.

**L'Homme, M.-C., B. Robichaud and P. Leroyer.** 2012. Encoding Collocations in DiCoInfo: From Formal to User-friendly Representations. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 211-236. Oxford: Oxford University Press.

**Park, Mi Sun.** 2013. Code-switching and Translanguaging: Potential Functions in Multilingual Classrooms. *Working Papers in TESOL & Applied Linguistics* 13(2): 50-52.

**Sinclair, J.** 2004. *Trust the Text Language, Corpus and Discourse.* London: Routledge.

**Smadja, F.** 1993. Retrieving Collocations from Texts: Xtract. *Computational Linguistics* 19(1): 143-177.

**Williams, G.** 1998. Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3(1): 151-171.

**Williams, G., C. Millon and A. Alonso.** 2012. *Growing Naturally: The DicSci Organic E-Advanced Learner's Dictionary of Verbs in Science*. Fjeld, Ruth Vatvedt and Julie Matilde Torjusen (Eds.). 2012. *Proceedings of the 15th Euralex International Congress, 7–11 August 2012*: 1008-1013. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

## Electronic term banks

*InterActive Terminology for Europe. The EU's Multilingual Term Base*. [Online] Available at: http://iate.europa.eu/. [Accessed 10 August 15].

*Le grand dictionnaire terminologique. Office québécois de la langue française.* [Online] Available at: http://www.granddictionnaire.com/. [Accessed 18 August 15].

*TERMIUM Plus®. The Government of Canada's Terminology and Linguistic Data Bank*. [Online] Available at: http://www.btb.termiumplus.gc.ca. [Accessed 30 August 15].

## Software

*SketchEngine.* Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý and V. Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography* (2014): 1-30. https://www.sketchengine.co.uk/.

*TlTerm.* TshwaneDJe. Human Language Technology. http://tshwanedje.com.

*WordSmith Tools.* Scott, M. 2012. WordSmith Tools version 6, Stroud: Lexical Analysis Software. http://www.lexically.net/wordsmith/.