Source Domain Verification using Corpus-based Tools

Kathleen Ahrens[1]

Menghan Jiang[2]

[1]*Department of English & Research Centre for Professional Communication in English, The Hong Kong Polytechnic University, Kowloon, Hong Kong*

[2]*Department of Chinese and Bilingual Studies & Research Centre for Professional Communication in English, The Hong Kong Polytechnic University, Kowloon, Hong Kong*

Corresponding author:

Kathleen Ahrens, Ph.D.

Department of English & Research Centre for Professional Communication in English

The Hong Kong Polytechnic University

Hung Hom, Kowloon

Hong Kong

Phone: 2766 7534

Email: kathleen.ahrens@polyu.edu.hk

Abstract

Source domain verification has not received as much attention as criteria for metaphor identification (Pragglejaz Group, 2007; Steen 2010) in the study of conceptual metaphor. In this paper, we provide a replicable approach to source domain verification which we hope will provide a foundation for new approaches to this important question. We adopt an empirical method extended from previous research that used corpus-based linguistic tools such as SUMO (Suggested Upper Merged Ontology), WordNet, collocational patterns and an online dictionary. We present a new, step-by-step procedure to verify which keywords may be categorized in the source domain of BUILDING, using data from the Corpus of Hong Kong Political Speeches which contains parsed Chinese-language speeches by Hong Kong Chief Executives of the Hong Kong Special Administrative Region (1997-2014). Following the verification of a number of keywords in the BUILDING source domain, we discuss how this method may be adapted for other source domains and languages and discuss its application to various areas of study within metaphor research as well as the current limitations of this approach.

Keywords: corpus linguistics, conceptual domain, Chinese, suggested-upper-merged-ontology, collocation patters

1. Introduction

Research in conceptual metaphor theory has traditionally been based on researcher intuition to identify whether or not an expression is used metaphorically or not. In the past decade, criteria for metaphor identification has been standardized to a great degree by relying on guidelines (Pragglejaz Group, 2007; Steen 2010) that are aided by dictionary entries so as to assist human judgement in verifying a potential metaphor in a given context. However, verifying what source domain a metaphor belongs to is has not been standardized the way metaphor identification has.

In an overview of corpus-based approaches to metaphor analysis, Stefanowitsch (2006) notes that previous scholars often begin their studies by selecting a potential source domain (i.e., a semantic domain or field that is known to play a role in metaphorical expressions), and then searching for individual lexical items from this domain. These lexical items are mostly selected based on manual selection. For example, by investigating four terms under the domain of TEMPERATURE, Deignan (1999a, 1999b, 2006) examined the syntactic (e.g., the Part-of-Speech of the lexical item), collocational and semantic (e.g., the semantic relation) patterning of linguistic metaphors, as well as the conceptual mapping between the source and target domain. In other studies, the lexical items were selected based on the keyword analysis of texts based on the topics under certain domain. For example, Partington (1998, 2003, 2006) generated the list of key items for the investigation of metaphors by conducting keyword analysis of texts based on different target domain topics (e.g., political discourse).

These previous studies mainly focused on a limited collection of lexemes under a certain source domain (either based on manual selection or keyword analysis). Therefore, Stefanowitsch (2006) suggested building a corpus annotated with semantic fields/domains (i.e. the source domain). He suggests that with this information annotated, we can specify a potential

source domain and search directly for all lexical items belong to that source domain. However, what is first needed is a systematic methodology for verifying what source domain a metaphorical keyword belongs to. He also pointed out that most studies categorize metaphors based on more-or-less explicit common sensical intuitions of the part of the scholars. This strategy may be problematic for cases that are not clear cut. An explicit source domain verification procedure can also help to alleviate these ambiguities.

Other approaches have been taken in Chinese to postulate the source domain between source and target domain pairings of conceptual metaphors by using two major databases: WordNet (1.6) and SUMO nodes (Chung, Huang and Ahrens (2003), Ahrens, Chung and Huang (2004), Chung, Ahrens and Huang (2005), Chung and Ahrens (2006), Huang, Chung and Ahrens (2007)).[1] They used WordNet relations and SUMO definitions to identify the relationships between metaphorical expressions and their corresponding ontological nodes. This approach has been shown to effectively reduce the manual work required for the verification of the source domain. However, a systematic approach that is potentially replicable over a variety of source domains or languages has not been undertaken. Moreover, in this study, due to the relatively large number of keywords, we found the identification of source domains using WordNet or SUMO tools alone was not enough. For example, some infrequent lexical items are not included in WordNet and SUMO, e.g., 架构 *jia4gou4* 'structure'. Hence, additional tools needed to be included to a more comprehensive methodology for source domain verification, including collocation patterns (Gong, Ahrens and Huang (2008), Chung (2009), Chung and Huang (2010)), as well as the use of an online dictionary.

[1] Shutova and Teufuel (2010) take a related approach in English using a subset of categories from the Master Metaphor List (http://araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf).

Therefore, in this paper we address the issue of source domain verification by proposing a series of steps that can be used to check hypothesized source domains with the assistance of corpus-linguistic tools: SUMO (Suggested Upper Merged Ontology), WordNet, an online dictionary and collocational patterns, along with a step-by-step procedure to utilize these tools. We demonstrate that verifying source domains in this way allows for an easy-to-use and replicable method of ascertaining whether a particular keyword belongs in a particular source domain or not. We suggest that this method is a useful tool for particular types of metaphor analyses with a number of commonly referenced source domains and may also be useful when drawing contrastive analyses regarding how speakers from different groups or over different time periods have used a particular source domain so as to gain insight into their stance or ideological viewpoint.

2. Source Domain Verification Procedure

Source domain verification involves first identifying potential keywords and then ascertaining if they belong in a hypothesized source domain or not. This may be done before or after metaphor identification occurs, as identifying potential metaphors is a separate procedure.[2] Because we are using a corpus that is three hundred thousand words, and because part of a future study is to see how political leaders use the BUILDING source domain in Hong Kong Policy Addresses, for the purposes of the current study we will verify whether or not the keywords are part of the BUILDING source domain prior to metaphor identification. Following the work of Charteris-Black (2004), among others, generating the potential keyword list was created by reading through a portion of the corpus carefully and identifying possible

[2] One advantage to ascertaining source domains before identifying metaphors is that it is then possible to contrast what concepts are mapped to a target domain and which ones are not in a given corpus. This may vary for corpora from different genre, such as medicine or politics. An advantage to ascertaining metaphors first is that there will then be fewer examples to analyze, since literal instances will already be ruled out.

metaphorical keywords as potentially belonging in the source domain of BUILDING (建筑 *jian4zhu4* 'building').[3]

Once we ascertained the keyword list, we then identified the language resources available in Chinese that are potentially useful for source domain verification. We selected four different language resources (1) Suggested Upper Merged Ontology (Niles and Peace, 2001), (2) WordNet (1.6) (Miller, 1995; Fellbaum, 1998), (3) an online Chinese dictionary (Handian, 2004), as well as (4) the Word Sketch Function in Sketch Engine (Kilgarriff, Huang, Rychly et al., 2005) and incorporated them into our decision-making process. In what follows we first explain the resources that we used and then we outline the steps involved in the verification procedure.

2.1 Corpora-based resources utilized for source domain verification

First, WordNet is a large-scale lexical knowledge base that was created at the Cognitive Science Laboratory of Princeton University in 1990, in which English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept (Fellbaum, 1998; Miller et al., 1990). In addition, WordNet also functions as a semantic network linking synsets with lexical semantic relations and is widely used in Natural Language Processing applications and linguistic research (Huang, Chang and Lee, 2004).

Second, SUMO (Suggested Upper Merged Ontology, http://ontology.teknowledge.com) is an upper ontology constructed by the IEEE Standard Upper Ontology Working Group and maintained at Teknowledge Corporation. SUMO and its domain ontologies form the largest formal public ontology in existence today (Niles and Pease, 2003). As of February 2003, the ontology contains more than 1000 terms and 4000 assertions. The purpose of SUMO is to be a

[3] If a previous researcher has postulated keywords to be in a particular source domain and these appear in the corpus, these may also be included in the verification process.

shared and inter-operable upper ontology (Niles and Pease 2001, Pease and Niles 2002, Sevcenko 2003). Since ontologies are formalized descriptions of the structure of knowledge bases, SUMO can also be viewed as a proposed representation of shared human knowledge, and thus is a good candidate for providing mapping information about the source domain (Ahrens, Chung and Huang, 2003; Chung, Huang and Ahrens, 2003; Ahrens, Chung and Huang, 2003, 2004)).

Both WordNet and SUMO were created for English; what we utilized in this study was a tool which integrated these two language resources for Chinese, which can be found at the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW—http://bow.ling.sinica.edu.tw/). Sinica BOW integrates WordNet, SUMO, and English-Chinese Translation Equivalents Database (ECTED), functioning both as an English-Chinese bilingual Wordnet and a bilingual lexical access to SUMO.[4] The goal of Sinica BOW is to give each linguistic form a rigorous conceptual location, and to clarify the relation between conceptual classification and linguistic instantiation, as well as to facilitate genuine cross-lingual access of knowledge (Huang et al., 2004). Sinica BOW allows versatile access and provides a combination of lexical semantic, and ontological information in a 2x2x2 query design, either in lexical lemmas or SUMO terms, and the query target can either be the WordNet content or the SUMO ontology (Huang et al., 2004). For example, in a WordNet search, the return includes an expandable list of the complete bilingual WordNet fields. The fields are listed under each sense and include: POS, synset, sense explanation, translation, and list of lexical semantic relations. In addition, the domain information, translation equivalents, and link to the

[4] The three above resources were originally linked in two pairs: the English synsets in WordNet were mapped to Chinese lexical equivalents by ECTED, and WordNet 1.6 was mapped to SUMO by Niles and Pease (2003). Thus, WordNet synsets were a mediating link for the integration work (Huang et al., 2004).

corresponding SUMO node are also presented and lead to the corresponding node in the domain taxonomy of the ontology to allow further exploration (Huang et al., 2004).

In order to verify the source domain of the keywords on our list, we first check the SUMO nodes of the keywords to ascertain whether they are good candidates for knowledge representation in the source domain. We then postulate the source domain of the keyword by checking the categories and definitions of the metaphorical keywords provided in WordNet as facilitated by the Sinica Bow interface, to see whether the most concrete meaning clearly aligns with a postulated source domain based on its WordNet sense or explanation. The proposed procedures will be tested and examples will be shown in detail in the following section.

Moreover, due to the relatively large number of keywords, we found the identification of source domains using WordNet or SUMO tools alone was not enough. For example, some lexical items are not included in WordNet and SUMO, e.g., 架构 *jia4gou4* 'structure'. In this case, we may turn to a third option: the word sense provided by an online dictionary may provide information for source domain verification (i.e. if the word sense provided by dictionary aligns with a postulated source domain). We select a free online Chinese dictionary for Chinese data—*Handian* (2004), which contains a number of authoritative Chinese dictionaries, including the *Advanced Chinese Modern Dictionary (1996)*.

Lastly, we note the usefulness of another analysis (Gong, Ahrens and Huang, 2008) which proposes that the conceptual domain of a word may also be ascertained by examining its collocates to verify findings based on WordNet, SUMO or a dictionary sense, or when none of these options have provided a definitive answer.[5] The tool to used here for collocations is the Word Sketch function in Sketch Engine, which processes a word's collocates and other

[5] Chung (2009) and Chung and Huang (2010) also have used collocational information when seeking to determine which source domains are related to a given target domain.

words in its surroundings. It summarizes the word's grammatical and collocational behavior and is sorted with the most typical collocations at the top (Kilgarriff et al., 2010), showing the searched word, its frequency, its collocates sorted into grammatical relations (e.g., objects, subjects, modifiers), the frequency of each collocate, and typicality score (Kilgarriff et al., 2010).[6]

2.2 The source domain verification procedure

The source domain verification procedure is illustrated in Figure 1. Only one of the four conditions needs to be fulfilled to verify a source domain.

The procedure runs as follows: After reading through the corpus and previous research to select keywords as postulated members of a particular source domain, the criteria for a keyword to be categorized in the source domain of BUILDING is determined. This is done by examining SUMO's nodes and deciding which conceptual nodes are related to the source domain of BUILDING. We selected the classes of "Stationary Artifact", "Building" and "Architecture", including any one of their subclasses (e.g., "Entertainment building", "Farm building", "Government building", "Library building", "Office building", etc.).[7] For verbal keywords, we decided that the class of 'Constructing' indicated that the keyword was in the source domain of BUILDING.[8] Thus, if a keyword has a conceptual node in any one of the above classes, it is considered as part of the source domain of BUILDING.

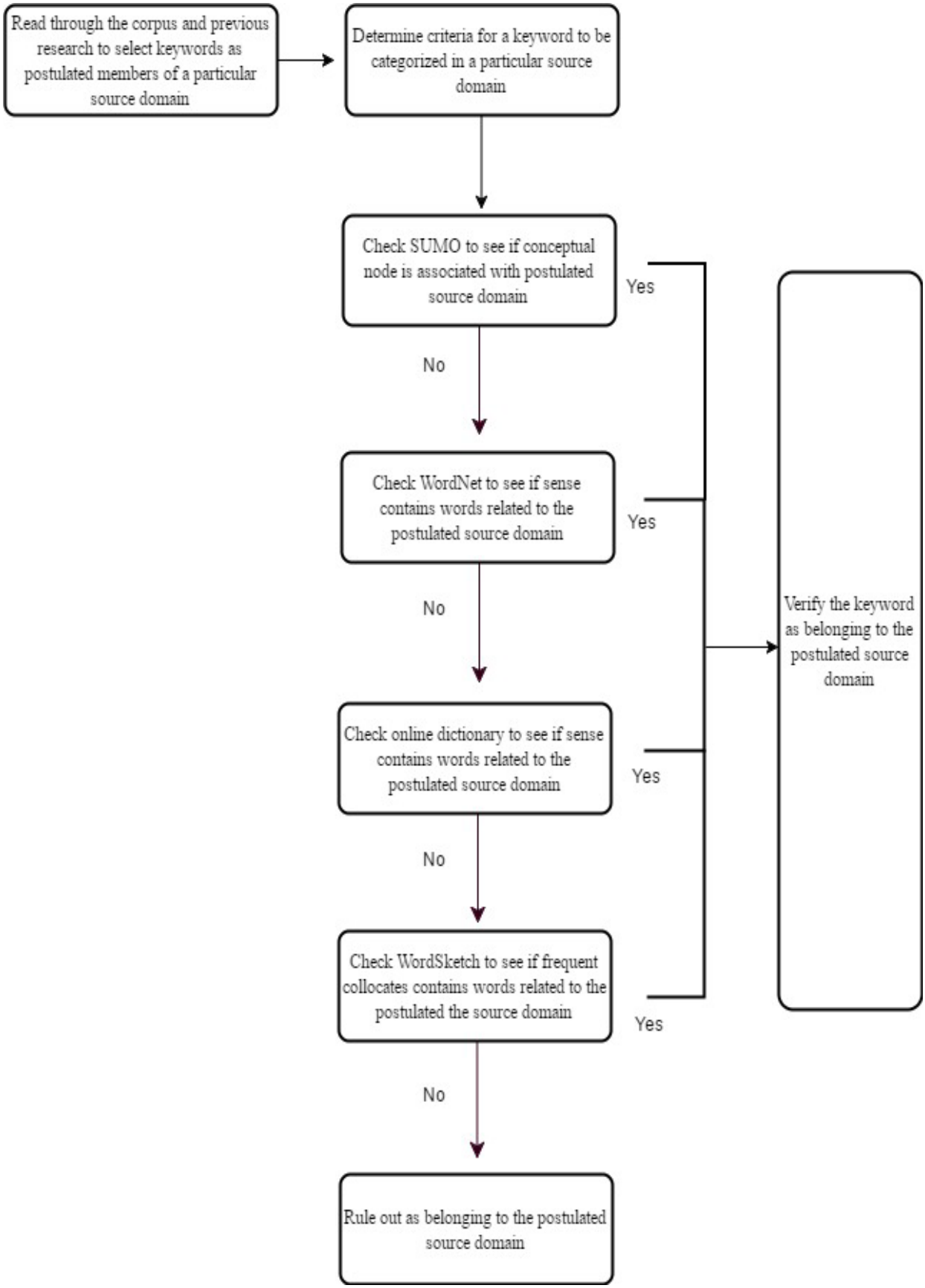
Figure 1. The source domain verification procedure

[6] Sketch Engine used a version of MI-Score modified to give greater weight to the frequency of the collocation. A very high score of the collocate means that there is little competition from other collocates because the node (i.e., the search word, the keyword) does not often combine with other collocates (Kilgarriff et al., 2014).

[7] SUMO defines a 'stationary artifact' as: 'an Artifact that has a fixed spatial location.' 'Class' in upper ontology is defined as abstract group, set, or collection of objects. Most instances of this Class are architectural works, e.g. the Eiffel Tower, the Great Pyramids, office towers, single-family houses, etc. The words 'fixed spatial location' and 'Architectural works' are terms that allow us to confirm the suggested source domain – 'BUILDING'.

[8] For a full taxonomy of the class of Building in SUMO, please see: http://sigma.ontologyportal.org:8080/sigma/Browse.jsp?lang=EnglishLanguage&flang=SUO-KIF&kb=SUMO&term=Building. To view the entire taxonomy of SUMO, please see: http://www.adampease.org/OP/images/SUMOclasses.gif.

SOURCE DOMAIN VERIFICATION



Next, if the SUMO nodes of the keyword do not clearly indicate the conceptual domain of the keyword belongs to BUILDING, we next try to verify the source domain by checking the categories and definitions of the keywords provided in WordNet. By searching the definitions of a keyword in WordNet in Sinica Bow using the Chinese-English look-up search engine, we locate a list of senses of the word (together with the explanation provided for the senses). The most concrete sense can be identified from this list. If the most concrete sense meets one of the following four conditions, we confirm the source domain as BUILDING.[9]

(1) Criteria for a keyword to be categorized in a particular source domain using WordNet, dictionary senses or collocates (using the source domain of BUILDING as an example)

a) The word sense and its explanation contain the word "building/house/architecture" (i.e., in Chinese this would be: 建筑物/楼房/房屋/房子/大厦/大楼 *jian4zhu4wu4/lou2fang2/fang2wu1/fang2zi0/da4sha4/da4lou2* 'building', etc.), as well as the subclasses of building including "office building", "government building", "residential building", "high rise", "factory building" etc.

b) The word sense and its explanation contain the word which refers to the components of a building, e.g., "balcony", "pillar"[10]

c) The word sense and its explanation contain the word which refers to different kinds of building (constructional engineering), including "bridge", "speedway" etc.

[9] These criteria will need to be developed independently for each source domain analyzed, with the understanding that identifying and then using these criteria allow for others to have a framework to verify these decisions, instead of relying solely on intuition alone.

[10] We included criteria b) and c) as they are similar to what can be found under the conceptual domain of 'stationary artifact' in SUMO.

d) The word sense and its explanation contain the word which refers to the act of building, e.g., "to build".[11]

Next, if neither WordNet nor SUMO contain enough information to decide, an on-line dictionary is referenced to check whether the word sense provided by the dictionary meets one of the four conditions shown above.

If WordNet, SUMO and the dictionary do not provide clear evidence of a semantic relationship between the metaphorical keyword and the hypothesized source domain (e.g., BUILDING), collocation searches for the keywords may be run using Chinese Sketch Engine (Kilgarriff, Huang, Rychly et al., 2005). We search for collocates of the keyword to check if BUILDING-related words found in the senses listed in (1) are frequently collocated with the potential keyword by using Chinese Sketch Engine.[12] In order to determine the cut-off point for collocating frequency, we use the notion of high saliency values. According to Chung and Huang (2010), the saliency values of the collocates may be separated into significant collocates and non-significant collocates using methods proposed in Chung et al. (2007), whereby a cut-off point for the significant collocates may be determined in terms of several different calculations. We follow the calculation of the 'mean of means', which is a threshold value that is computed based on the mean of a group of means of saliency values (cf. Chung, 2007, 2009;

[11] We included criteria b) and c) as they are similar to what can be found under the conceptual domain of 'Constructing' in SUMO.

[12] Chinese Sketch Engine is based on the Chinese Gigaword Corpus, which can be found at http://wordsketch.ling.sinica.edu.tw/. Registration is required to access Chinese Sketch Engine.

Chung and Huang, 2010).[13] An example showing how the cut-off point is calculated will be explained in Section 3.

Lastly, if neither the SUMO nodes, the WordNet senses and explanations, the word senses in dictionary nor the syntactic collocation has indicated that the word is semantically related to the suggested source domain, then we exclude this potential keyword from the source domain category.

3.0 Testing the Source Domain Verification Procedure using Hong Kong Chief Executives' Corpus

In this section, we run the proposed procedure on a set of Chinese keywords hypothesized to be part of the source domain of BUILDING and found in a small political corpus.[14]

3.1 Corpus

The corpus utilized for this study is the Hong Kong Chief Executives Corpus (1997-2014), which is one of the sub-corpora of the HKBU Corpus of Political Speeches (http://digital.lib.hkbu.edu.hk/corpus/index.php) (Ahrens, 2015). The Hong Kong Chief Executives corpus is comprised of Hong Kong Policy Address delivered annually by Chief Executive of the Hong Kong Special Administrative Region to the Hong Kong Legislative Council.[15] The address includes a summary of the past work of the Hong Kong government and an introduction of its policy in the coming year. This report is released through various

[13] The formula for mean of means is shown below.

$$\frac{mean_1(Saliency_1, Saliency_2) + \ldots + Mean_n(Saliency_{(n-2)}, Saliency_{(n-1)}, Saliency_{(n)})}{n-1}$$

[14] Note that utilizing a specific corpus allows the researcher to read through a portion of the text and identify potential keywords specific to that corpus that may be missed if solely relying on intuition or previous work.

[15] In Chinese this report is known as 施政报告.

media outlets and most people see it as a useful way of predicting what the Hong Kong politicians will focus on in the coming year. There are three speakers involved in this corpus: Tung Chee-hwa, Donald Tsang Yam-kuen, and Leung Chung-ying. All speeches were written in both English and Chinese. As the purpose of this paper is to evaluate the usefulness of the source domain verification procedure, we will focus on the Chinese version of Hong Kong Policy Addresses. Table 1 presents the details of the Corpus.

Table 1. Hong Kong Chief Executives' corpus of political speeches

Speakers	**Year**	**Word count**
Tung Chee-hwa	1997-2005	169,654
Donald Tsang Yam-kuen	2006-2012	144,965
Leung Chun-ying	2013-2014	53,320
Subtotal		367,939

3.2 Identifying potential keywords and criteria for inclusion in the source domain

One linguist trained in metaphorical analysis read through the first and the last political speech for each of the three Hong Kong politicians, which in total contains 367,939 words and and identified 39 metaphorical keywords as potentially belonging in the source domain of BUILDING (建筑 *jian4zhu4* 'building'), as presented in Table 2. A group of four linguistically trained speakers of Chinese also determined the criteria, based on SUMO and WordNet, that needed to be met for a keyword to be considered as part of the source domain of BUILDING.

Table 2. Keywords hypothesized to be in the source domain of BUILDING

N	V	N/V
板块 *ban3kuai4* 'board'	缔造 *di4zao4* 'build'	平衡 *ping2heng2* 'balance'
布局 *bu4ju2* 'layout'	奠基 *dian4ji1* 'lay the foundation'	稳定 *wen3ding4* 'stable'
层 *ceng* 'floor'	封顶 *feng1ding3* 'seal roof'	支持 *zhi1chi2* 'support'
底层 *di3ceng2* 'ground floor'	巩固 *gong3gu4* 'strengthen'	
基层 *ji1ceng2* 'base'	缓冲 *huan3chong1* 'buffer'	
工程 *gong1cheng2* '(constructional) engineering'	加强 *jia1qiang2* 'strengthen'	
基础 *ji1chu3* 'foundation'	建构 *jian4gou4* 'construct'	
基石 *ji1shi2* 'cornerstone'	建立 *jian4li4* 'build'	
架构 *jia4gou4* 'structure'	建设 *jian4she4* 'build'	
结构 *jie2gou4* 'structure'	扩充 *kuo4chong1* 'extend'	
空间 *kong1jian1* 'space'	扩大 *kuo4da4* 'enlarge'	
框架 *kuang1jia4* 'frame'	扩展 *kuo4zhan3* 'expand'	
平台 *ping2tai2* 'platform'	扩张 *kuo4zhang1* 'expand'	
楼 *lou2* 'building'	强化 *qiang2hua4* 'strengthen'	
枢纽 *shu1niu3* 'hinge'	调整 *tiao2zheng3* 'adjust'	
支柱 *zhi1zhu4* 'pillar'	拓宽 *tuo4kuan1* 'broaden'	
中枢 *zhong1shu1* 'pivot'	拓展 *tuo4zhan3* 'expand'	
中心 *zhong1xin1* 'center'	牢固 *lao2gu4* 'solid'	

3.3 Checking SUMO nodes

Next, the SUMO nodes of these 39 words were checked on the Sinica Bow interface http://bow.ling.sinica.edu.tw/. There were four keywords that clearly aligned with a postulated

source domain (i.e. BUILDING) based on their SUMO nodes of 'Stationary artifact,' 'Building,' and 'Constructing' as discussed above: 底层 *di3ceng2*, 楼 *lou2* 'building', 强化 *qiang2hua4* 'strengthen' and 加强 *jia1qiang2* 'strengthen' (Table 3). Hence, these four keywords may be ascertained to be part of the source domain of BUILDING.

Table 3. SUMO nodes for BUILDING metaphorical keywords

Metaphorical Keyword	SUMO nodes
底层 *di3ceng2* 'ground floor'	Stationary artifact (固定人造物)
楼 *lou2* 'building'	Building (建筑物)
强化 *qiang2hua4* 'strengthen'	Constructing (建筑)
加强 *jia1qiang2* 'strengthen'	Constructing (建筑)

3.4 Checking WordNet Senses

Next, for the remaining 35 keywords on our list, the categories and definitions of the keywords provided in WordNet were also checked. In this step, a total of eleven words were confirmed to be in the BUILDING domain, including 基层 *ji1ceng2* 'base', 层 *ceng2* 'floor', 建立 *jian4li4* 'build', 基础 *ji1chu3* 'foundation', 基石 *ji1shi2* 'cornerstone', 平台 *ping2tai2*

'platform', 扩展 *kuo4zhan3* 'expand', 缔造 *di4zao4* 'build', 建设 *jian4she4* 'build', 奠基 *dian4ji1* 'lay the foundation', 工程 *gong1cheng2* '(constructional) engineering'.

For example, for the keyword 基石 *ji1shi2* 'cornerstone' was found to be 'Artifact' when we checked its SUMO node (column 4 in Table 4), which is superordinate to 'stationary artifact', and thus less specific. Thus, we needed to move to the next step and check its WordNet sense. By searching the definitions of 基石 *ji1shi2* 'cornerstone' in WordNet in Sinica Bow using the Chinese-English look-up search engine, we located a list of senses of the word. The most concrete sense (i.e., the more concrete meaning that was mapped from the source domain) was identified from the list. In this case, the sense 'cornerstone' (column 2 in Table 4) was selected as the most concrete sense and 'BUILDING' is chosen as the suggested source domain for 基石 *ji1shi2* 'cornerstone', based on the explanation provided for the sense (column 3 in Table 4).

Table 4. WordNet-SUMO definition of 基石 *ji1shi2* 'cornerstone'

Metaphorical Keyword	WordNet Sense	Explanations	SUMO nodes
基石 *ji1shi2* 'cornerstone'	cornerstone	a stone laid at a ceremony to mark the founding of a new **building**	Artifact (人造物)

3.5 Checking Dictionary Senses

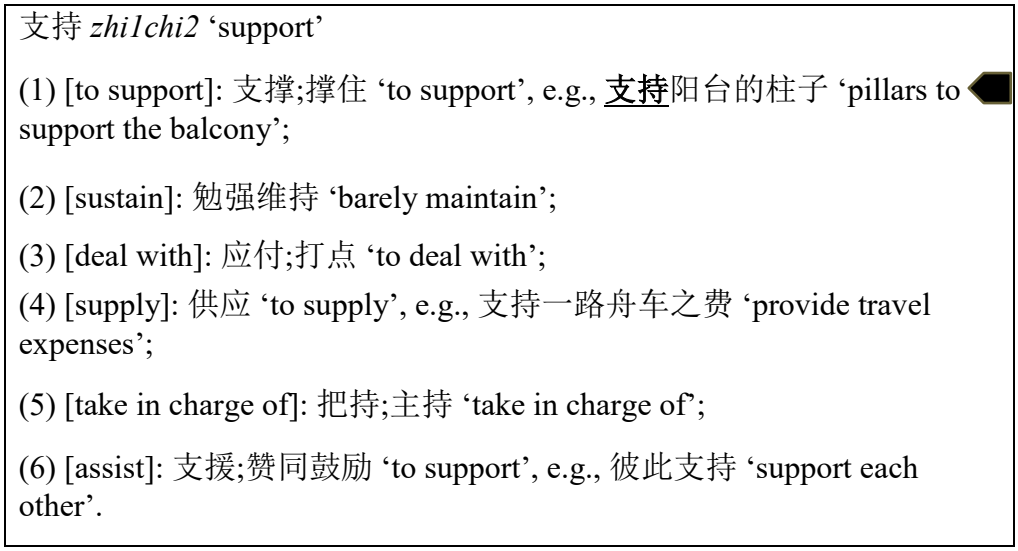
For the remaining 24 keywords, neither WordNet nor SUMO contained enough information to reach a decision. In those cases, the *Handian* online Chinese Dictionary was referenced (https://www.zdic.net/). For instance, for the keyword 支持 *zhi1chi2* 'support', WordNet-SUMO does not provide specific evidence about which source domain it relates to, as shown in Table 5, where the SUMO node given is 'Process.'

Table 5. WordNet-SUMO search of 支持 *zhi1chi2* 'support'

Metaphorical Keyword	WordNet Sense	Explanation	SUMO nodes
支持 *zhi1chi2* 'support'	support	supply the force or power for the functioning of	Process

We then checked the word senses provided in the *Handian* online Chinese dictionary, where 支持 *zhi1chi2* 'support' is defined as 支撑, 撑住, 例如支持阳台的柱子 'to support, such as the pillars to support the balcony' (as shown in point 1 of Figure 2). Since the word 'pillars' and 'balcony' are included in the definition, as noted *a priori* in 1(b), we conclude that BUILDING is the source domain for 支持 *zhi1chi2* 'support'.

Figure 2. Word Senses in *Handian* Online Dictionary



The word sense in *Handian* provides evidence for in total 8 words, including 框架 *kuang1jia4* ‘frame’, 架构 *jia4gou4* ‘structure’, 拓宽 *tuo4kuan1* ‘broaden’, 封顶 *feng1ding3* ‘seal roof’, 牢固 *lao2gu4* ‘solid’, 结构 *jie2gou4* ‘structure’, 巩固 *gong3gu4* ‘strengthen’ and 支持 *zhi1chi2* ‘support’, which allow us to confirm their proposed source domain as BUILDING.

3.6 Checking Collocational Patterns

For the remaining 16 words, neither Wordnet, SUMO or *Handian* provide clear evidence of a semantic relationship between the metaphorical keyword and the hypothesized source domain (i.e. BUILDING), so collocation searches for the keywords are run using Chinese Sketch Engine (Kilgarriff, Huang, Rychly et al., 2005) in order to verify the suggested source domain BUILDING. Three keywords are confirmed as being under BUILDING metaphor in this step are 支柱 *zhi1zhu4* ‘pillar’, 建构 *jian4gou4* ‘construct’ and 稳定 *wen3ding4* ‘stable’.

We take 支柱 *zhi1zhu4* ‘pillar’ as an example. We used Chinese Sketch Engine to search for collocates of 支柱 *zhi1zhu4* ‘pillar’ to see if it frequently co-occurs with 建筑 *jian4zhu4* ‘building’. Figure 3 shows the collocates of 支柱 *zhi1zhu4* ‘pillar’ in terms of various grammatical relations to the keyword. The first column in the figure includes all the collocates, the second column is the frequency of the collocates, and the third column is the saliency value for that collocation pair. As we mentioned in Section 2, we ascertained the cut-off points (i.e. mean of means) for each grammatical relation to find the significant collocates for each keyword. For example, for the saliency list below in Figure 3, the first mean is the mean of the saliency values of the first (4.36) and the second collocates (4.29); and the second mean is the mean of saliency values of the first three collocates: collocate one (4.36), two (4.29), and three (3.62), i.e., we add a new collocate each time. When all means have been calculated for all collocates, an overall mean is obtained from all these means (this is the ‘mean of means’). In this case, the threshold value for the possessors of 支柱 *zhi1zhu4* is 2.716, and for the noun modifiers of 支柱 *zhi1zhu4* is 4.633. As we can see from the figure, among the significant collocates of 支柱 *zhi1zhu4*, there is a BUILDING related word 骑楼 *qi2lou2* ‘arcade building’ which shows a saliency value of 6.63, which is greater than 4.633. Following the criteria set up in (1a) above along with the criteria for high saliency, we can conclude indicating that 建筑 *jian4zhu4* ‘building’ is one of the top collocations of the keyword 支柱 *zhi1zhu4* ‘pillar’.

Figure 3. Results of Significant Collocation for 支柱 *zhi1zhu4* ‘pillar’ (Chinese_giga_trd freq = 2040)

Collocates	**Freq**	**Saliency Value**	**Collocates**	**Freq**	**Saliency Value**
Possessor		**33.09**	**N_Modifier**		**22.3**
背后 *bei4hou4* ‘at the back’	4	4.36	吊杆 *diao4gan1* ‘Derrick boom’	3	7．53
中华民族 *zhong1hua2min2zu2* ‘Chinese nation’	3	4.29	底层 *di3ceng2* ‘ground-floor’	3	6.19
胜选 *xuan3sheng4* ‘win’	3	3.62	骑楼 *qi2lou2* ‘arcade building’	6	6.13
家庭 *jia1ting2* ‘family’	19	3.09	精神 *jing1shen2* ‘spirit’	122	5.23
			江 *jiang1* ‘river’	9	5.05

3.7 Excluding keywords from the source domain

If none of the language resource provided clear information to indicate the potential keyword has semantic relationship with BUILDING, then we excluded this potential keyword from the list. There were 13 keywords selected by the annotator as a potential metaphorical keyword related to BUILDING based on expert’s intuition. However, neither the WordNet senses and explanations, the SUMO nodes, the word senses in dictionary, nor the syntactic collocation indicated the word is semantically related to BUILDING. Two of them were very abstract and cannot be asserted as under any metaphorical source domain: 缓冲 *huan3chong1* ‘buffer’, and 平衡 *ping2heng2* ‘balance’. There were 11 keywords that may potentially belong to other source domains, including SPACE (e.g., 中心 *zhong1xin1* ‘center’, 空间 *kong1jian1* ‘space’, 扩充 *kuo4chong1* ‘extend’, 扩张 *kuo4zhang1* ‘expand’, 扩大 *kuo4da4* ‘enlarge’, and 扩展 *kuo4zhan3* ‘expand’), GAME (e.g., 布局 *bu4ju2* ‘layout’), MACHINE (e.g., 调整 *tiao2zheng3* ‘adjust’), EARTH (e.g., 板块 *ban3kuai4* ‘board’) and BODY (e.g., 中枢 *zhong1shu1* ‘pivot’).

3.8 Discussion

In conclusion, after two linguists worked through the source domain verification procedure for the 39 potential keywords, 26 of them were found to have evidence for being associated with the source domain of BUILDING (Table 6). Among these 26 keywords, 4 decisions were based on the SUMO nodes and definitions, 11 decisions were based on the WordNet sense and explanations, 8 were based on definitions in the on-line dictionary, and 3 decision were based on collocation evidence (the shaded cells indicate at what step each decision was made). In addition, we can see from Table 6 that all four options were needed in order to ensure robust coverage for inclusion in this source domain, as ruling out any given step would result in the exclusion of some keywords as other steps would not necessarily provide the needed information.

Table 6. Confirmed keywords in different language resources ("Y" indicates evidence can be found to verify the source domain of BUILDING and the shaded cells indicate the decision point)

BUILDING Keywords	SUMO	WordNet	Handian	Word Sketch
楼 *lou2* 'building'	Y	Y	Y	Y
底层 *di3ceng2* 'ground floor'	Y	Y	Y	Y
加强 *jia1qiang2* 'strengthen'	Y			
强化 *qiang2hua4* 'strengthen' (V)	Y			
层 *ceng* 'floor'		Y	Y	Y
平台 *ping2tai2* 'platform'		Y	Y	Y
工程 *gong1cheng2* '(constructional) engineering'		Y	Y	Y
基层 *ji1ceng2* 'base'		Y	Y	
基础 *ji1chu3* 'foundation'		Y	Y	
基石 *ji1shi2* 'cornerstone'		Y	Y	
建设 *jian4she4* 'build'		Y		Y
奠基 *dian4ji1* 'lay the foundation'		Y		Y
建立 *jian4li4* 'build'		Y		
扩展 *kuo4zhan3* 'expand'		Y		
缔造 *di4zao4* 'build'		Y		
拓宽 *tuo4kuan1* 'broaden'			Y	Y
封顶 *feng1ding3* 'seal roof'			Y	Y
牢固 *lao2gu4* 'solid'			Y	Y
结构 *jie2gou4* 'structure'			Y	Y
支柱 *zhi1zhu4* 'pillar'			Y	
巩固 *gong3gu4* 'strengthen'			Y	
架构 *jia4gou4* 'structure'			Y	
框架 *kuang1jia4* 'frame'			Y	
支柱 *zhi1zhu4* 'pillar'				Y
建构 *jian4gou4* 'construct'				Y
稳定 *wen3ding4* 'stable'				Y

Since all four steps are needed for maximum inclusion, the ordering of the steps we have proposed for the Source Verification Procedure reflects the fact that the SUMO verification step is the simplest to ascertain, followed by WordNet, dictionary senses and then collocations.

4. Conclusion

In this paper we proposed a method to verify which postulated keywords may be categorized within the source domain of BUILDING with the assistance of four corpus-based resources: SUMO (Suggested Upper Merged Ontology), WordNet, collocational patterns and an online dictionary. This method has the advantage of providing both flexibility and accountability to researchers working in various aspects of metaphor theory as it provides guidelines for how to undertake source domain determination and can be specified in a paper's methods. The specification would need to state the SUMO classes and sub-classes that were considered to be associated with this source domain as well as the lexical items that appear in the WordNet or dictionary senses that indicate association with this source domain, as in (1) above. The flexibility comes from the ability for researchers to make different choices regarding the SUMO classes, sub-classes, or the lexical items in WordNet or a dictionary that indicate association with the source domain. The accountability comes from being able to report this information clearly so that other researchers could replicate the analysis or question why certain choices were made and offer alternate specifications.

In addition to using this method to analyze source domains, this process may also be used when identifying target domains. For example, a target word concept, such as 'economy' may be searched in a corpus and then potential metaphor identification could occur, using either MIP (Pragglejaz Group, 2007) or MIPVU (Steen, 2010). If 'economy' has been used

metaphorically, the word or phrase in the sentence may then be identified as a potential keyword for a given source domain, after which the source domain verification process could occur. Furthermore, it could be explored as to whether SUMO may aid in metaphor identification, as its ontology classifies each sense of a word as either an entity that is either 'physical' or 'abstract' allowing for a researcher to ascertain that a word has a more concrete sense. Using SUMO may also help researchers in clarifying the levels of metaphor question that Kovesces raises in his (2017) paper. His proposal is that that there are four levels of schematicity, with images schemas as the most schematic, followed by domains, frames, and then mental spaces (which is the least schematic). BUILDING is a domain in this system and frames further elaborate aspects of those domains. It remains to be seen if the ontological structure of SUMO aids not only in source domain identification, but also may help with elaborating aspects of the frames within each domain.

Moreover, simply looking at source domains has potential implications as well, as this line of research may contrast the use of the source domains across genres or gender or time periods (i.e., comparing the British-appointed Hong Kong Governors use of a particular source domain from 1984 to 1996 with the Hong Kong Chief Executives who led Hong Kong after British Colonial rule ended in 1997). In addition, with this procedure and the bilingual tools available, it is also possible to contrast what source domains are used in the English versions of the Policy Addresses with the Chinese versions of the Policy Addresses.[16]

Of course, further work is needed to ascertain if this type of ontologically-based source domain verification works well for some source domains, such as BUILDING, or JOURNEY, but

[16] In addition to the bilingual tools used in this study, language generation templates for SUMO in Hindi, Italian, German and Czech can be found here: http://www.adampease.org/OP/. Information on Wordnets in other languages can be found here: http://globalwordnet.org/resources/wordnets-in-the-world/ and information on how to use SketchEngine in over ninety languages can be found here: https://www.sketchengine.eu/corpora-and-languages/.

not others, such as LIGHT/DARKNESS. Further research utilizing SUMO, WordNet and associated corpus-based information, including collocations, will give a clearer understanding as to which source domains are amenable to this particular type of analysis. In addition, it may be the case that collocations (or SUMO or WordNet or dictionary definitions) are especially useful for verifying particular source domains.

In sum, the source domain verification procedure is a new set of procedures that provides researchers with a principled, yet flexible, set of guidelines with which to determine if a postulated word belongs in a particular source domain. The analysis involves ascertaining the criteria to be followed using corpus-based tools and then checking these criteria for each word that is postulated to be in that source domain, allowing for greater research validity and replicability in metaphor studies.

Acknowledgements

The first author would like to thank the University Grants Council of Hong Kong for supporting this research (General Research Fund # 12400014). Both authors would like to thank Winnie Hui-heng Zeng, Jessie Shijie Zhang, Joanna Zhuoan Chen, Ivy Wing-Shan Chan, and Leslie Chen Tong for their assistance in building the corpus and in data extraction and analysis. Responsibility for any errors remains with the authors.

References

*Advanced Chinese Modern Dictionary 高级汉语词典(*Wang Tongyi eds). (1996). Hainan Press.

Ahrens, K. (2015). *Corpus of Political Speeches*. Hong Kong Baptist University Library Retrieved from http://digital.lib.hkbu.edu.hk/corpus/.

Ahrens, K., Chung, S. F., and Huang, C. R. (2004). From lexical semantics to conceptual metaphors: Mapping principle verification with wordnet and sumo. *In Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW)*, Singapore. pp. 99-106.

Ahrens, K., Chung, S. F., and Huang, C. R. (2003). Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. In: Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14. Association for Computational Linguistics, p. 36-42.

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Berlin: Springer.

Chung, S. F. (2009). *A corpus-driven approach to source domain determination*. Taipei: Institute of Linguistics, Academia Sinica.

Chung, S. F. (2007). *A corpus-driven approach to source domain determination*. Ph.D. Dissertation, Graduate Institute of Linguistics, National Taiwan University.

Chung, S. F., and Ahrens, K. (2006). Source Domain Determination: WordNet-SUMO and Collocation. *In Proceedings of the 2nd International Conference of the German Cognitive Linguistics Association,* München, Germany. pp. 1-4.

Chung, S. F. and Huang, C.-R. (2010). Using collocations to establish the source domain of conceptual metaphors. *Journal of Chinese Linguistics.* 38(2): 183-223.

Chung, S. F., Huang, C.R. and Ahrens, K. (2003). Economy is a Transportation-Device: contrastive representation of source domain knowledge in English and Chinese. *In International Conference on Natural Language Processing and Knowledge Engineering Proceedings*, Beijing, China. pp. 790-796.

Chung, S. F., Ahrens, K., and Huang, C. R. (2005). Source domains as concept domains in metaphorical expressions. *International Journal of Computational Linguistics & Chinese Language Processing, Special Issue on Selected Papers from CLSW-5, 10*(4), 553-570.

Chung, S. F., Ahrens, K., Cheng, C.-P., Huang, C.R., and Šimon P. (2007). Computing thresholds of linguistic saliency. In the *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 126-135.

Deignan, A. (2006). The grammar of linguistic metaphors. In Stefanowitsch and Gries (Eds.) *Trends in Linguistics Studies and Monographs, 171*, 106. Berlin: Mouton De Gruyter.

Deignan, A. (1999a). Corpus-based research into metaphor. In Cameron and Low (Eds.) *Researching and applying metaphor* (pp. 177-199). Cambridge: Cambridge University Press.

Deignan, A. (1999b). Metaphorical polysemy and paradigmatic relations: A corpus study. *Word, 50*(3), 319-338.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gong, S. P., Ahrens, K., and Huang, C. R. (2008). Chinese word sketch and mapping principles: A corpus-based study of conceptual metaphors using the BUILDING source domain. *International Journal of Computer Processing of Languages*, *21*(01), 3-17.

Handian (2004). *Handian Online Dictionary*. Retrieved from https://www.zdic.net/.

Huang, C. R., Chang, R. Y., and Lee, H. P. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. *In The International Conference on Language Resources and Evaluation,* Lisbon, Portugal. pp. 1553-1557.

Huang, C. R., Chung, S. F., and Ahrens, K. (2007). An ontology-based exploration of knowledge systems for metaphor. In Kishore, Rajiv, Ram Ramesh, and Raj Sharman (Eds.), *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*. Volume 14. Berlin: Springer. pp. 489-517.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. (2014). The Sketch Engine: ten years on. In Lexicography 1(1): 7–36. DOI: 10.1007/s40607014-0009-9. ISSN 2197-4292.

Kilgarriff, Adam, Vojtěch Kovář, Simon Krek, Irena Srdanovič, and Carole Tiberius. (2010). A quantitative evaluation of word sketches. *Proceedings of the 14th EURALEX International Congress*: 372-79, 2010.

Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. (2005). Chinese word sketches. *In the Proceedings of the 4th Asialex Conference*, Singapore.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM (Association for Computing Machinery)*, *38*(11), 39-41.

Miller, G. A., et al. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, 3(4): 235-244.

Niles, I., and Pease, A. (2003). Mapping WordNet to the SUMO Ontology. Teknowledge Technical Report.

Pease, A. and Niles, I. (2002). IEEE Standard Upper Ontology: A Progress Report. Knowledge Engineering Review, Special Issue on Ontology and Agents, Volume 17.

Niles, I., and Pease, A. (2001). Towards a standard upper ontology. *In Proceedings of the International Conference on Formal Ontology in Information Systems.* Ogunquit, ME, USA. pp. 2-9.

Sevcenko, M. (2003). Online Presentation of an Upper Ontology. In Proceedings of Znalosti 2003, Ostrava, Czech Republic, February

Partington, A. (2006). Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work. In Stefanowitsch and Gries (Eds.) *Trends in Linguistics Studies and Monographs, 171*, 267. Berlin: Mouton De Gruyter.

Partington, A. (2003). *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. Abingdon, Oxon: Routledge.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching* (Vol. 2). Amsterdam: John Benjamins Publishing.

Pragglejaz Group (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol, 22*(1), 1-39.

Shutova, Ekaterina, and Simone Teufel. (2010). Metaphor Corpus Annotated for Source-Target Domain Mappings. In *The International Conference on Language Resources and Evaluation.* Malta. pp. 3255-3262.

Steen, G. (Ed.). (2010). *A method for linguistic metaphor identification: From MIP to MIPVU (Vol. 14)*. Amsterdam: John Benjamins Publishing.

Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. In Stefanowitsch and Gries (Eds.) *Trends in Linguistics Studies and Monographs*, 171, 1-16. Berlin: Mouton De Gruyter.