Using a Corpus of English and Chinese Political Speeches for Metaphor Analysis

Kathleen Ahrens[1], Huiheng Zeng[2], Shun-han Rebekah Wong[3]

[1,2]Department of English, The Hong Kong Polytechnic University
11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong
[3]HKBU Library, Hong Kong Baptist University
Kowloon Tong, Kowloon, Hong Kong
kathleen.ahrens@polyu.edu.hk, winnie.zeng@connect.polyu.hk, rebekahw@hkbu.edu.hk

Abstract

In this article, we present details of our corpus of political speeches and introduce using the corpus for metaphor analysis in political discourse. Although specialized corpora on a variety of topics are now easily available, online political corpora available for public use are scarce. The database our research team has developed contains more than six million English and Chinese political speeches and is currently available free online. Researchers in many fields are able to use the multiple search functions on the website for their specific research purposes. In particular, the corpus is useful for researchers focusing on political speeches and conceptual metaphor analyses. From the perspective of metaphor study, we have taken advantage of several functions to facilitate the corpus-based metaphor analyses. In short, this database enriches the current bilingual resources and contributes to the evaluation of political language by linguists and political scientists.

Keywords: corpus, political speeches, metaphor

1. Introduction

The rising prominence of corpus-driven approaches in linguistic studies has sparked a growing interest in applying a corpus-based approach to metaphor analysis in the political domain (Charteris-Black, 2004, 2009; Ahrens, 2009; Deignan, 1999, 2005; Semino, 2006, 2008; Deignan and Semino, 2010). The corpora used in these studies are usually compiled manually by authors, i.e. Charteris-Black (2009) collected data of British parliamentary debates from online versions of Hansard, while Koller and Semino (2009) assembled a corpus of interviews and speeches by Germany chancellors from the official websites of the German government. However, these corpora have not been released and made available for public search or free use. Our project proposes the creation of a corpus entitled 'The HKBU Corpus of Political Speeches' which is a large-scale political database with English and Chinese data and featured search functions; it is convenient for corpus-based research and also available for online free use.

Two major design features of the HKBU Corpus of Political Speeches are as follows: 1. The minimal interface of the database website is designed for users to operate easily. Similar to Sinica Corpus 4.0 (Chen et al., 1996), the search functions of our database such as the keyword-in-context (KWIC) search and collocation search can retrieve the lexical frequency and collocation lists which facilitate a corpus-based approach to linguistic analyses. Our database is especially useful for metaphor analysts who are able to search keywords in the source domains and the target domains of a particular metaphor and extract all the data involving the keywords searched into Excel files for further analyses. In addition, the collocation search and left one sort or right one sort functions can contribute to the determination of specific source domains or target domains of the keywords. Examples of studies that have used corpora to aid in metaphor analyses are Ahrens and Chang (forthcoming) who studied the role of lexical frequency and collocation in postulating conceptual models and Ahrens, Chung and Huang (2004) and Chung, Ahrens and Huang (2004a, b) who studied the source domain determination

and verified the Mapping Principles in conceptual metaphors.

The corpus itself also contains bilingual versions of political speeches for Hong Kong Policy Addresses which provides a valuable resource for researchers focusing on translation. It also contains English and Chinese (traditional and simplified) political corpora with various speech types (i.e. presidential speeches, political debates, policy addresses, reports of government work). The speeches are delivered by politicians in different political positions (i.e. Presidents, Governors, Chief Executives and Premiers) and are from People's Republic of China, Hong Kong, Taiwan and the United States.

The corpus presented in this paper provides an intuitive way for political analysts, linguists and general users to access and use the political corpora as it provides for a simple interface that allows for in-depth and fine-grained analysis of these texts via KWIC search and collocation. In addition, we hope that the analysis of conceptual metaphors in English and Chinese will also allow a greater understanding of the rhetorical patterns in political speeches so as to understand the ideologies of politicians in areas from China, Hong Kong, Taiwan and the United States. It will also provide insight into how linguistic patterns drawn from corpora help to understand and build metaphor models. In the following sections, we will introduce our database in detail and present how it is applied in facilitating metaphor research.

2. Corpus of Political Speeches

2.1 Dataset creation

The HKBU Corpus of Political Speech (http://digital.lib.hkbu.edu.hk/corpus/) is an online archive of political speeches developed by our research team together with the Hong Kong Baptist University library. In total, 6,269,359 words of political speeches delivered by politicians from China, Hong Kong, Taiwan and the United States were collected for this online database. It currently contains four collections of speeches: the English Corpus

of U.S. Presidential Speeches (1789-2015), including six different types of sub-corpora; the English & Chinese Corpus of Policy Addresses by Hong Kong Governors (1984-1996) and Hong Kong Chief Executives (1997-2014); the Chinese Corpus of Speeches given on New Year's days and Double Tenth days by Taiwan Presidents (1978-2014) and the Chinese Corpus of Report on the Work of the Government by Premiers of the People's Republic of China (1984-2013). Details of the four collections are listed in Table 1.

US Corpus of Political Speeches (4,429,976 words)
Inaugural Addresses (1789-2013) – *in English*
Annual Messages to Congress on the State of the Union (1790-2014) – *in English*
National Political Party Platforms (1944-2012) – *in English*
Presidential Nomination Acceptance Speeches (1960-2012) – *in English*
Presidential Candidates Debates (1960-2012) – *in English*
Saturday Radio Addresses (1982-2014) – *in English*
HK Corpus of Political Speeches (1,079,712 words)
Policy Address by HK Governors (1984-1996) – *in English*
Policy Address by HK Chief Executives (1997-2014) – *in English*
Policy Address by HK Governors (1984-1996) – *in Chinese*
Policy Address by HK Chief Executives (1997-2014) – *in Chinese*
Taiwan Corpus of Political Speeches (169,649 words)
Speeches Given on New Year's days and Double Tenth days by Presidents in Chinese (1978-2014) – *in Chinese*
P.R.C. Corpus of Political Speeches (590,022 words)
Report on the Work of the Government by P.R.C. Premiers in Chinese (1984-2013) – *in Chinese*

Table 1: Corpus of English & Chinese Political Speeches

2.2 Chinese corpora annotation

For the Chinese Corpora of Hong Kong, Taiwan and PRC political speeches, we further annotated each corpus with part-of-speech tagging (Figure 1) by using The Stanford Natural Language Processing Software (SNLPG, 2015), Stanford Word Segmenter 3.7 for Chinese words segmentation and Stanford POS Tagger 3.7 for part-of-speech tagging.

Figure 1. POS tagging in Chinese corpus

To improve the tagging accuracy, two researchers in our team who are both native Chinese speakers with a linguistic background took six months to further check all the tagged texts word by word based on Stanford segmentation (Fei, 2000a) and part-of-speech tagging (Fei, 2000b). We found that the automatic word segmentation and POS tagging on the simplified Chinese is more accurate than the traditional Chinese. The accuracy for both word segmentation and POS tagging in the PRC corpus with simplified Chinese characters was approximately 85%. It was approximately 60% in Hong Kong and Taiwan Corpora which both use traditional Chinese characters. After two linguistically-trained native speakers of Chinese manually checked the corpus, the problematic taggings were revised and the tagged Chinese corpora provide a reliable Chinese database with a wide range of syntactically tagged texts. This allows researchers to conduct analyses of the syntactic structures, grammatical features and syntactic distribution in political discourse. It is especially useful to metaphor researchers focusing on Chinese political discourse. For example, the following sentence has been extracted from the database.

Example

高等/JJ 教育/NN 体制/NN 改革/NN 和/CC 结构/NN 调整/NN 的/DEG 步伐/NN 加快/VV 。/PU (朱镕基, PRC, 1999)
Gao1deng3/JJ jiao4yu4/NN ti3zhi4/NN gai3ge2/NN he2/CC jie2gou4/NN tiao2zheng3/NN de0/DEC bu4fa2/NN jia1kuai4/VV 。/PU (Zhu1 Rong2ji1, PRC, 1999)
Higher/JJ_education/NN_system/NN_reform/NN_and/CC _structure/NN_adjustment/NN_of/DEG_pace/NN_speed up/VV ./PU (Zhu Rongji, PRC, 1999)
'The pace of the high education system reform and structural adjustment has been speeded up. (Zhu Rongji, PRC, 1999)'

The sentence has been segmented into separate lexical units which is usually considered to be the first step of the metaphor identification procedure (Pragglejaz Group, 2007; Steen et al., 2010). Metaphor analysts can work directly on the segmented sentence to check the basic/contextual meanings of each lexical unit in order to locate keywords that have been metaphorically used. In this example, three words 结构/NN *jie2gou4*/NN 'structure', 调整/NN *tiao2zheng3*/NN 'adjust' and 步伐/NN *bu4fa2*/NN 'pace' are identified as the metaphorical keywords. In addition, the Chinese POS tagging provides syntactic structure information contributing to the determination of word senses and collocations.

2.3 Features and functions of the database

The corpus has a web-based concordance feature, which performs a number of functions for corpus searches in untagged texts (for English corpora) and part-of-speech tagged texts (for Chinese corpora). Firstly, the KWIC searches are for the number of occurrence, graphical representation and the full-text request available (Figure 2).

Figure 2. No. of Occurrence

The searches are able to be further refined by choosing different sources, speakers of the speeches and time frames (Figure 3). Collocation search is available for the top 20 words searches and the search can be specified by choosing different positions of the words (up to 10) before or after the keywords (Figure 4).
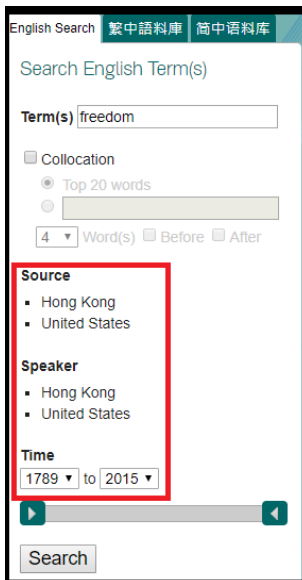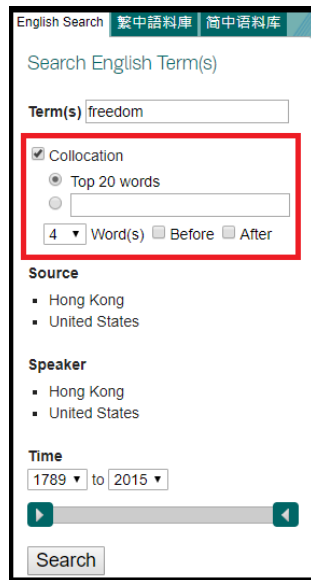

Figure 3. Search filters   Figure 4. Collocation search

For example, if we search the collocations of 'freedom' (one word before 'freedom') in the Hong Kong Pre-1997 corpus, we will get the number of occurrence and the graphical representation of all the collocates (Figure 5).
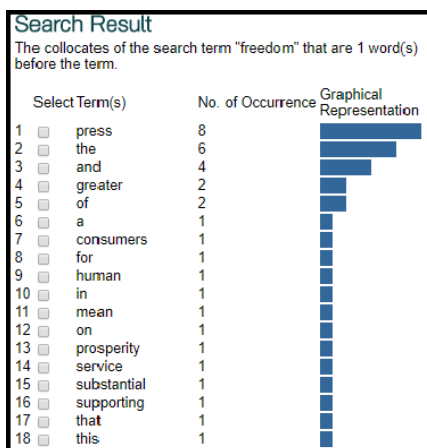
Figure 5. Collocation search results

Then if we select the most frequent collocate 'press', all the sentences involving 'press' will be retrieved and presented. The search results can be further sorted by positions of the search words (from L5 to R5), years, regions, types and speakers of target speeches (Figure 6).

Figure 6. Search filters

In addition, the system allows users to export all selected sentences into Excel files for further analysis (Figure 7), which is especially useful to metaphor researchers who need to look at surrounding context to determine of a word or phrase is being used literally or metaphorically.

Figure 7. Extracted sentences in Excel files

Links to the original official sources have been provided for users who want to read the full-text. Detailed instructions of how to use the search functions for both Chinese corpus and English corpus can be found in the User Manual at http://digital.lib.hkbu.edu.hk/corpus/help.php.

3. Corpora-based Evaluation of Political Metaphor

Metaphor research examines what concrete concepts in one conceptual domain are mapped to abstract concepts in another domain. Copora-based conceptual metaphor research has shown that corpora-based analyses may provide linguistic evidence of metaphor variations by comparing metaphor use within the same source or target domains (Charteris-Black, 2009; Ahrens, 2009, 2011; Lu and Ahrens, 2008; Ahrens and Lee, 2009; Semino and Koller, 2009). We believe that a cross-linguistic analysis of

conceptual metaphor based on abundant corpus data will allow researcher to evaluate the degree to which metaphors occur cross-linguistically and provide linguistic evidence of the mapping principles between the source and the target domains (Ahrens, 2010). These types of studies will contribute to the investigation of existing metaphor theories.

For example, using corpora from our English database (the State of Union Addresses and Presidential Radio Addresses of US Presidents 1981-2006), Ahrens, (2011) has investigated the lexical frequency patterns related to two metaphor models (i.e. the Strict Father model and the Nurturant Patent model) and found support for Lakoff's (1996, 2002) hypothesis that Democrats and Republicans view the world differently. Evidence from the Reagan and Clinton data show that they used metaphors based on value paradigms in two different metaphor models, respectively. The findings in that paper led to a better understanding of the underlying conceptual worldview of the political leaders in the Democratic and Republican parties and pointed out the limitations of a purely intuition-based approach.

In addition, metaphor researchers may take advantage of the search functions in the Corpus of Political Speeches to evaluate metaphor in different corpora at multiple levels (i.e. analysis at cross-regional, cross-linguistically, cross-speakers levels and from diachronic angles). Since the data in our corpus can be searched for within a particular time period, this enables analysis of diachronic language change and the top lexemes for each speaker from different regions allows the comparison of lexical usages between different Chinese corpora or between Chinese and English corpora. Frequency listed word counts provide additional linguistic information for more in-depth and finer grained analysis of lexical and metaphor use. We apply approaches of top-down analysis to first search all the keywords in our database and extract target texts to Excel files in order to examine the selected keywords. We then normalize the frequency of metaphorical keywords within a particular source or target domain for the comparison of metaphor usage and their underlying political ideologies.

For example, Lu and Ahrens (2008) found that Kuomintang Presidents in Taiwan used BUILDING metaphors to instill a Chinese ideology. However, the president from the Democratic Progressive Party preferred not to use BUILDING metaphors, and instead used FARMLAND metaphors to emphasize Taiwan's agricultural background and political independence. In addition, Kuomintang Presidents used BUILDING metaphors in ways that differ from US Presidents. The Kuomintang Presidents used retrospective BUILDING metaphors to emphasize the past history of China, while US Presidents used BUILDING metaphors to emphasize creating a particular type of structure (i.e., economic, educational and political) for future generations. In Hong Kong English corpus, metaphor analysis of the target domains with the source domains of JOURNEY and BUILDING showed that the Hong Kong Chief Executives tend to use more JOURNEY than BUILDING metaphors and the target domain refers primarily to Hong Kong's future. In addition, when Chief Executives talk about Hong Kong's future, the future often collocates with positive evaluations (Ahrens, 2016a, b).

Our on-going metaphor analyses (Ahrens and Zeng, 2017a, b) focus on the comparison of source domain types related to the target domains of EDUCATION and DEMOCRACY in both Chinese corpus (HK, PRC and TW corpus) and English corpus (HK corpus). Figure 8 presents examples of Chinese EDUCATION metaphor analyses in PRC corpus. Column A indicates the time frames of the speeches; column B shows the regions of the speeches delivered; the names of the speakers are indicated in column C; column D and F are the speech texts extracted; column E indicates the target domain keyword 教育 jiao4yu4 'education'; in column G, we input the source domains manually identified and in column H we input the metaphorical keywords that we used to determine the source domains listed in column G. We hope that in the future these manual analyses will provide a full set of data that can be used to aid future metaphor research.

Figure 8. Examples of EDUCATION metaphor analyses

Figure 9 illustrates one example of the normalized ratio results related to the comparisons of the source domains (PRODUCT & BUILDING) of EDUCATION metaphors between Hong Kong Chief Executives and PRC Premiers. The results show that Hong Kong Chief Executives conceptualize education as PRODUCT more frequently than PRC Premiers while both groups use the concept of BUILDING with similar frequencies. Within Hong Kong Chief Executives, Tung Chee-hwa focuses more on the concept of education as PRODUCT and Donald Tsang uses more BUILDING metaphors. PRC premiers understand education more as BUILDING than PRODUCT, among whom Li Peng and Wen Jiabao apply the two source domains more frequently than Zhao Ziyang and Zhu Rongji.
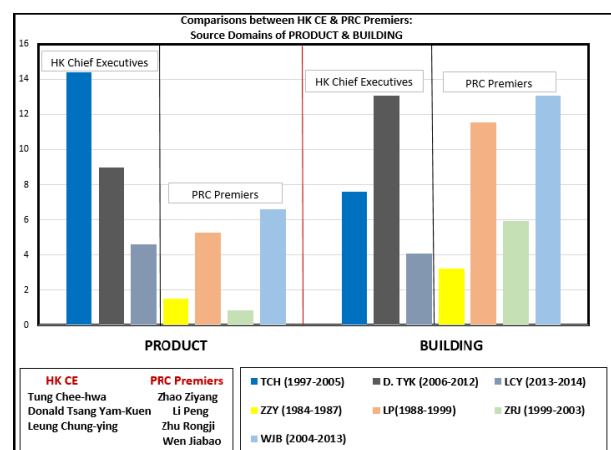

Figure 9. Comparisons of source domains between Hong Kong Chief Executives and PRC Premiers

Evaluating metaphor source domain variations in the Chinese corpora in our database will also allow for contrastive analysis of Chinese languages over the past three decades, which will lead to a deeper understanding of the degree of universality that languages share with the same cultural backgrounds and the degree of difference that languages reflect in different political ideologies. All these will allow greater understanding of the rhetorical patterns of persuasion in political speeches across time.

4. Conclusion and Future Work

In this paper, we presented the HKBU Corpus of Political Speeches and described the information about the dataset as well as about the Chinese data annotation and the featured search functions of the database. We also introduced corpus-based metaphor studies in political discourse by using the search functions and corpora from this database.

For the further development of the corpus, we hope to: 1) enlarge our corpus by continuing to update the current corpora and adding new speeches to compile new corpora, such as to include political speeches by Hong Kong senior officials, PRC chairmen or speeches by politicians from Macau; 2) use Stanford part-of-speech tagger to annotate all the English corpora with POS tagging in order to align with the POS tagged Chinese corpora; 3) try to build a small corpus with tagged metaphors.

5. Acknowledgements

6. Bibliographical References

Ahrens, K. (Ed.). (2009). *Politics, gender and conceptual metaphors*. Basingstoke and New York: Palgrave-MacMillan.

Ahrens, K. (2010). Mapping principles for conceptual metaphors. In L. Cameron, A. Deignan, G. Low & Z. Todd (Eds.), *Researching and applying metaphor in the real world)*. Amsterdam: John Benjamins, pp. 185-207.

Ahrens, K. (2011). Examining conceptual metaphor models through lexical frequency patterns: A case study of U.S. presidential speeches. In H. J. Schmid (Ed.), *Windows to the mind. Series: Applications of cognitive linguistics.* Berlin: Mouton De Gruyter, pp. 167-184.

Ahrens, K. (2016a). English conceptualizations in a Chinese context: BUILDING metaphors in Hong Kong policy addresses. Presented at International Conference on Boundary Crossing and Bridge Building. National Taiwan University, Taipei, Taiwan, 17 June, 2016.

Ahrens, K. (2016b). Contrasting BUILDING metaphors in the Corpus of Political Speeches. Presented at ICAME (International Computer Archive of Modern & Medieval English) 37. Chinese University of Hong Kong, Hong Kong, 25-29 May 2016.

Ahrens, K. and Chang, P. Y. C. (Forthcoming). A Corpus-based approach to conceptual metaphor modeling: The role of lexical frequency and collocation. In C. R. Huang, S. K. Hsieh, & Z. F. Sui (Eds), *Chinese Language Resources: Data collection, linguistic analysis, annotation, and language Processing.* Springer.

Ahrens, K. and Lee, S. (2009). Gender versus politics: When conceptual models collide in the U.S. Senate. In K. Ahrens (Ed.), *Politics, gender, and conceptual metaphors.* Basingstoke and New York: Palgrave-MacMillan, pp. 62-82.

Ahrens, K. and Zeng, H. H. (2017a). Metaphors of Education in China and Hong Kong. Presented at Look Both Ways: Narrative & Metaphor in Education - Conference 2017. Vrije University, Amsterdam, 30 March - 1 April 2017.

Ahrens, K. and Zeng, H. H. (2017b). The Conceptualization of Democracy in Hong Kong Policy Addresses. Presented at The third VEIP (Varieties of English in the Indo-Pacific) meeting. Linguistics Macquarie University, Australia, 16-17 February 2017.

Ahrens, K., Chung, S. F., and Huang, C. R. (2004). From lexical semantics to conceptual metaphors: Mapping principle verification with WordNet and SUMO. In Ji Donghong, Lua Kim Teng, and Wang Hui, editors, Recent Advancement in Chinese Lexical Semantics. Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5), pages 99-106. Singapore: COLIPS.

Charteris-Black, J. (2004). Corpus approaches to critical metaphor analysis. Basingstoke: Palgrave Macmillan.

Charteris-Black, J. (2009). Metaphor and gender in British parliamentary debates. In A. Kathleen (Ed.), *Politics, gender and conceptual metaphors.* Basingstoke and New York: Palgrave-MacMillan, pp. 139-165.

Chen, K. J., Huang, C. R., Chang, L. P., and Hsu, H. L. (1996). SINICA CORPUS: Design methodology for balanced corpora. In Byung-Soo Park and Jong-Bok, editors, Proceedings of Language, Information and Computation (PACLIC 11), pages 167-176. Seoul, Korea.

Chung, S. F., Ahrens, K., and Huang, C. R. (2004a). Using WordNet and SUMO to determine source domains of conceptual metaphors. In Ji Donghong, Lua Kim Teng, and Wang Hui, editors, Recent Advancement in Chinese Lexical Semantics. Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5), pages 91-98. Singapore: COLIPS.

Chung, S. F., Ahrens, K., and Huang, C. R. (2004b). RECESSION: Defining source domains through WordNet and SUMO. Proceedings of the Linguistic Society of Korea (LSK) International Conference. Volume II: General Sessions, pages 43-52. The Linguistic Society of Korea (LSK) and Yonsei Institute of Language and Information Science (ILIS). Seoul.

Deignan, A. (1999). Corpus-based research into metaphor. In A. Deigan, G. Low & Z. Todd (Eds.), *Researching and applying metaphor in the real world*. Cambridge: Cambridge University Press, pp. 177-200.

Deignan, A. (2005). *Metaphor and corpus linguistics (Vol. 6).* Amsterdam: John Benjamins Publishing.

Deignan, A. & Semino, E. (2010) Corpus techniques for metaphor analysis. In: L. Cameron & R. Maslen (Eds.), *Metaphor analysis: Research practice in applied*

linguistics, social sciences and the humanities. London: Equinox, pp. 161-179.

Fei, X. (2000a). The segmentation guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*. Retrieved from: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1038&context=ircs_reports

Fei, X. (2000b). The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*. Retrieved from: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1039&context=ircs_reports

Koller, V., and Semino, E. (2009). Metaphor, politics and gender: a case study from Germany. In A. Kathleen (Ed.), *Politics, gender, and conceptual metaphors*. Basingstoke: Palgrave Macmillan, pp. 9-35.

Lakoff, G. (1996). *Moral politics: What conservatives know that liberals don't*. Chicago: University of Chicago Press.

Lakoff, G. (2002). *Moral politics: How liberals and conservatives think*. Chicago: The University of Chicago Press.

Lu, L. W. L. and Ahrens, K. (2008). Ideological influences on BUILDING metaphors in Taiwanese presidential speeches. *Discourse and Society, 19*(3): 383-408.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*(1): 1-39.

Semino, E. (2006). A corpus-based study of metaphors for speech activity in British English. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy.* Amsterdam: John Benjamins Publishing Company, pp. 35-60.

Semino, E. (2008). *Metaphor in discourse*. Cambridge: Cambridge University Press.

Semino, E. and Koller, V. (2009). Metaphor, politics and gender: a case study from Italy. In A. Kathleen (Ed.), *Politics, gender and conceptual metaphors.* Basingstoke: Palgrave-MacMillan, pp. 36-61.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., and Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

The Stanford Natural Language Processing Group. (2015). Stanford Log-linear Part-Of-Speech Tagger (Version3.6.0) [Software]. Available from htttp://nlp.stanford.edu/software/tagger.html