CHAPTER 4

Locating metaphor candidates in specialized corpora using raw frequency and keyword lists

Gill Philip
Università di Macerata, Italy

This chapter explains one method that can be used to extract linguistic metaphors from a specialized corpus of Italian political speeches, using statistically-based measures incorporated into most standard corpus query software – in this case, WordSmith Tools (Scott 2004). This method can be used alone or in combination with existing manual or semi-manual analyses. While software has been developed for the automatic extraction of metaphors in English, minority languages, including Italian, lack tools for semantic annotation and probability measures that underlie such applications. The method presented in this chapter is intended for users who have no access to lemmatizers, semantic taggers, etc., and/or are working with under-resourced languages, for which no such tools are generally available.

Keywords: Italian, metaphor detection, under-resourced languages

1. Introduction

When using corpora in metaphor studies, the question always arises as to how metaphors are to be located. Corpora are designed to facilitate the extraction of forms, but metaphors are not formally different from other words: they are merely words that are being used with a metaphorical sense. While it has been demonstrated that metaphorically used words tend to collocate differently from their non-metaphorical counterparts (see Deignan 2005, Deignan & Potter 2004, Partington 2003), unless the distinct collocational patterns have been identified in advance, the researcher cannot take advantage of the differences in formulating his or her search queries. A further complication arises when the metaphorical meanings are not common enough to appear in a reference corpus, making it impossible to analyse their patternings.

This chapter aims to offer researchers a method for locating metaphors in corpus data that does not rely on prior investigation of word forms and their collocates, nor

does it require extensive reading and annotation of the corpus texts. For reasons that are detailed below (Section 4), the success of the approach is dependent on the corpus being homogeneous as far as its topical content (subject matter) is concerned; it thus responds to the growing interest in domain-specific language in corpus studies and increasing use of corpora in metaphor research. It also responds to the pressing need for methods that are independent of dedicated software applications, which are not language-specific, and which can be used by the individual researcher using a PC concordance package to analyse an un-annotated corpus.

2. What is a metaphor?

Some metaphors are more metaphorical than others. This has less to do with the metaphorically used word than it does with the way that an individual interprets that word. In the approach taken here, some metaphors are treated as metaphorically motivated terminology rather than as metaphors proper, and it is necessary to explain the reasoning that lies behind this decision.

In a strict definition of metaphor, any word that is used to mean something different from its main, or core, sense, is being used metaphorically.[1] Determining precisely what the core sense of a word is not without its problems. If by *core* we mean *literal*, then all non-literal uses must, by default, be metaphorical. While this allows for a clear-cut differentiation between literal and non-literal to be made, it is not necessarily the most practical measure to adopt, not least because several definitions of literal coexist (see Lakoff 1986, Gibbs et. al 1993). If, however, we choose to take *salience* as the benchmark against which to measure metaphoricity, the distinction between figurative and non-figurative blurs markedly. Salient meanings are those that are the "coded meanings foremost on our mind due to conventionality, frequency, familiarity, or prototypicality" (Giora 2003: 10). A salient sense need not be literal, as demonstrated by the morass of dead metaphors, in which the metaphorical sense, by force of linguistic habit, is used without regard to its figurative nature, but rather as if it were simply a homonym of its literal counterpart. In brief, a word can be used metaphorically without it necessarily being perceived consciously as *a metaphor*.

While the literature abounds with distinctions of metaphorical vitality (whether alive or dead, or somewhere in between), less attention has been paid to context- and usage- dependent classifications of metaphor. Metaphor is in the eye of the beholder, as it were, and there are some parameters that affect the perception of metaphoricity. One is the semantic parameter, which distinguishes between dead metaphors and their live counterparts, though this is not discussed here (see the classifications in Black 1993, Goatly 1997, and discussion of these in Deignan 2005; on dead metaphor in

1. See, for example, the metaphor identification procedure outlined in Pragglejaz Group (2007); also see Section 3 and Kaal and Dorst (this volume).

particular, the reader is referred to Lakoff 1987). The dimensions that are considered here, because they are directly relevant to the metaphor location method to be outlined, are related to matters of familiarity operating on three interrelated planes: the pragmatic, the textual, and the personal.

The more conventional a metaphorically-used word or expression is, the less metaphorical it seems, because its conventionality acts as a buffer, weakening the elaboration of the metaphorical entailment and thus dulling the imagery invoked by the metaphor. This is true of the semantic dimension as well as the pragmatic one, but of particular importance to pragmatic meaning is the fact that a conventional metaphor is not used so much for the conceptual or visual correspondences that it sets up (as their freshness has waned), but rather for the pragmatic force that has come to be associated with that particular expression. That force is not inherent in the metaphorically used word, but is a result of its use in conventionalized collocational patternings. Similar to the concept of the "metaphoreme"[2] (Cameron & Deignan 2006), which is familiar to metaphor scholars, the pragmatic force associated with a metaphor in its lexical context is well-established in corpus linguistics, where it is associated with the term "semantic prosody" (Sinclair 1996)[3], the most abstract and intangible element of the "extended unit of meaning" (ibid). The conventionalized patternings associated with the metaphorically used word comprise lexical and grammatical features – respectively, in Sinclairean terms, "collocates" (words that repeatedly co-occur with the expression) and "colligates" (grammatical classes that repeatedly co-occur with the expression, including syntactic and textual positioning in addition to collocation of items belonging to the same word class: see especially Hoey 2005). Variety in collocates belonging to a particular semantic or lexical set lead to the identification of the "semantic preference" (Sinclair 1996). The semantic prosody, however, can be glossed as "what is really being conveyed by the use of this chunk", and is a complex combination of semantic meaning, attitude and evaluation, and the circumstantial and contextual (extralinguistic) factors surrounding its use. So an established metaphor conveys not only an established (semantic) meaning, but also an established set of associative meanings and an established pragmatic force (see Philip 2009b). In contrast, a novel metaphor, which by definition is unconventional and therefore has not yet built up its own set of typical patternings, does not occur as part of a unit of meaning but, instead, as a free-standing element: it can thus only rely on the power of word meaning, and its pragmatic value is gleaned from extemporaneous features alone, not from established use.

Notions of conventionality are not absolute, however, and one of the most noticeable ways in which conventionality can be misinterpreted is in specialized discourses.

---

2. Editors' note: The term "metaphoreme" is "posited as a bundle of lexico-grammatical, cognitive, semantic, pragmatic and affective features around a phrase that has metaphorical meaning, and that has emerged over time from discourse" (Cameron 2010: 336). See also Gibbs, this volume.

3. See also Aksan and Aksan, this volume.

Terminological, restricted, or domain-specific meanings attract patternings that are conventional in a particular discourse, but are not conventional in general language. Members of the discourse community adhere to its norms of usage: domain-specific vocabulary is acquired, used and interpreted in the form that is conventionalized for that discourse. Outsiders to that discourse perceive such discourse-conventional forms in a different light, however, effectively over-interpreting their meaning because the unit of meaning in operation is not conventional in the discourses with which they are familiar. As a result, the outsider is far more likely to notice that a word is being used metaphorically than a discourse community member is: while 'growth' and 'flow' are used metaphorically in economics discourse to talk about the generation of wealth and the exchange of money respectively, an economist uses these words as terms, not as metaphors, while linguists repeatedly fall into the trap of considering them as metaphors whose entailments require investigation. The stance taken in this chapter is the following: metaphorically motivated terminology is used as terminology, not for its metaphorical value. It is thus excluded from consideration in a study of metaphor, because although its meaning is metaphorical if judged with reference to an earlier, *original*, sense, it is not used metaphorically. Partington expresses the concept succinctly: a dead metaphor is "an item which has ceased to collocate, in a particular genre, with the set of items it collocated with in its earlier sense" (2003: 210).

This brings us to the final parameter, that of the individual's experience of the language. While conventionality is a fact pertaining to the language and its community of speakers, familiarity lies with the individual. A conventional metaphor is not necessarily familiar to all speakers; this is especially true for language learners, but it is equally the case for native speakers who have simply not come into contact with the metaphor in question. Irrespective of the reasons why a conventional metaphor is unfamiliar, the result is to opt for a salient-meaning-first strategy in determining its meaning (Giora 1999). Thus the word-meaning value of the metaphor is effective, with the pragmatic force of the semantic prosody remaining largely inaccessible. I say *largely inaccessible*, because it is not necessarily absent: when a conventional metaphor is being used, this is for its pragmatic meaning, as expressed through the semantic prosody, not for the surface wording (but see Philip 2011, Chapters 6 and 7 on creative uses and variation). Speakers are quite able to infer the intended pragmatic meaning from other contextual cues, and perceive the mismatch between the words used and the (presumed) intended meaning (ibid., Chapter 4).

While conventional use of metaphor is identifiable in corpus data, the same cannot be said for familiar use of metaphor. Although this aspect of comprehension should always be taken into consideration, it cannot be investigated empirically using corpus data because it deals with the personal rather than the collective, and it therefore does not affect the findings of corpus-based research. Conventional language use is realized in corpus data in consistent collocational patternings, making it possible for the researcher to distinguish those words that are used metaphorically from those that are metaphorical but used conventionally, such as metaphorically-motivated terminology.

This matter is particularly important when dealing with unfamiliar specialized domains, as it enables the researcher to filter out his or her own perceptions of metaphoricity (see Section 3), and instead measure it according to the norms of the discourse community in and for which the language concerned has been produced.

3. Locating metaphors in text

The identification of metaphors in text is a time-consuming and labour-intensive business. There are various possible approaches that can be adopted, but all necessarily involve the close reading of the text(s) concerned, sometimes by more than one researcher, in order to identify metaphorically-used words as defined by the agreed classification criteria. The metaphor identification procedure (MIP), described in Pragglejaz Group (2007: 3), uses the following criteria to identify metaphorically used words:

> If the lexical unit has a more basic current-contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.

By following this procedure, both content and structure words may be classed as metaphorical, while it is common elsewhere to disregard apparently non-literal uses of structure words (for example, the prepositions 'in' and 'on') and focus exclusively on content words. This stringent method, while ensuring replicability, does not respond to the needs of all researchers, not least because it treats each word (i.e. a string of characters surrounded by white space) in isolation from those around it, and thus cannot account for multi-word units and meaning expressed over word boundaries. For this reason, some researchers prefer to use modified versions of this procedure, (see, for example, Low, Littlemore & Koester 2008), while others still use the criterion of incongruity to identify metaphors (Cameron 2003, Charteris-Black 2004). In some cases, researchers do not specify their criteria for classifying uses as metaphorical (e.g. Partington 2003).

Reading texts in a linear fashion from beginning to end has its advantages and disadvantages: on the one hand, the sequential identification procedure makes it difficult for metaphors to slip through the net as each word is considered in turn, and borderline cases can be checked one at a time against the classification criteria. However, as with any activity requiring human concentration and judgement, errors, omissions and misclassifications may arise, even if the work is being double-checked by another researcher. The study reported by Pragglejaz Group (2007) illustrates clearly both how metaphor identification can be carried out, and how even well-trained experts may differ in their judgements when following the same classification criteria.

If this procedure is carried out on all the texts comprising a corpus, that corpus can then be tagged for its metaphorical content and queries performed on these tags. However, it is not common for entire corpora to be tagged for their metaphorical

content. Scholars whose work focuses on metaphor in discourse and who use corpora for this purpose (Charteris-Black 2004, 2005, Partington 2003, Koller & Semino 2009, Semino & Koller 2009) inevitably encounter problems due to the sheer volume of data to be analysed. Corpora are generally too large for manual analysis to be considered feasible, so the problem of identifying metaphors tends to be overcome by combining such close reading with concordancing. The identification procedure in this case is generally performed in two stages, one manual, the other automated. The first stage involves close, word-by-word reading of a sample of the corpus texts, and then the findings obtained from this analysis are used as the basis for the corpus analysis proper, in which concordances of the identified words and expressions are called up (see Berber Sardinha this volume). In addition to the actual words found in the manual analysis, the researcher may choose to include others that s/he believes are likely to occur, for example, synonyms and semantically related forms, as well as co-inflected forms of the identified lemmas, e.g. including the plural form of a word that has been identified in the singular as being metaphorical.

> Charteris-Black reports the following procedure:
> My approach to metaphor identification has two stages: the first requires a close reading of a sample of texts with the aim of identifying candidate metaphors. [...] These candidate metaphors were then examined in relation to the criteria for the definition of metaphor specified in Chapter 1. It will be recalled that these were the presence of incongruity or semantic tension – either at linguistic, pragmatic, or cognitive levels – even if this shift occurred some time before and has since become conventionalized. Those that did not satisfy this criterion were excluded from further analysis. Words that are commonly used with a metaphoric sense are then classified as metaphor keywords and it is possible to measure the presence of such keywords quantitatively in the corpus. The second stage is a further qualitative phase in which corpus contexts are examined to determine whether each use of a key-word is metaphoric or literal. (2004: 35)

Partington (2003: 198–210) takes a very different approach to metaphor identification, eschewing any manual analysis whatsoever. He extracts keywords from his data (a genre-specific corpus) then computes n-gram clusters (consecutive strings of 4–5 words) featuring these keywords. Different meanings are characterized by distinct phraseological patternings, so metaphorical uses can be distinguished from non-metaphorical ones (ibid: 199). Koller & Semino (2009) and Semino & Koller (2009) use a combination of these two approaches. In the first instance, they manually analyse a core sample of the data – around 25% – following the MIP procedure (Pragglejaz Group 2007) to identify the metaphors used. They then compare the metaphors to a keyword list to see whether any of them were key in the corpora studied. Concordancing is carried out on both the keywords and the metaphoric expressions, and an n-gram tool is used to extract the phraseological patternings associated with the metaphorical words and expressions (Koller & Semino 2009).

Even if partial manual analysis plus corpus querying makes it possible for metaphors to be studied in large data sets, it is more problematic than manual analysis alone. Recurrent metaphors, or groups of lexically or semantically related metaphors, inevitably predominate in such analyses, because corpus searches for the identified words (and any related ones that the researcher wishes to include) will result in the retrieval of further instances of those words in the remainder of the data. In contrast, any metaphors (whether one-off or recurrent) not found in the part of the corpus that was processed manually remain invisible. The metaphors are present in the data, but are hidden.

4. Locating metaphor candidates

4.1 Background

How can the problem of retrieving metaphors be solved? For the researcher working on English language data, there are ways of getting round it. English is probably the best resourced of all languages as far as text processing tools are concerned, with lemmatising and part-of-speech and semantic tagging easily available even for researchers working with corpora that they have compiled themselves. The situation for researchers working with other languages can be quite different. For example, Italian – the language used to illustrate the method in this chapter – has no national corpus, and the only true general reference corpus is somewhat limited in its functionality; annotation tools such as lemmatizers and part of speech (POS) taggers exist but are not made available outside the research teams that have developed them, meaning that outsiders cannot benefit from them. The individual researcher who has compiled a corpus can only rely on the data in its raw form, and the statistical calculations that come as an integral part of many PC concordance packages. It is with these resources in mind that a method for locating metaphor has been developed.

4.2 Preliminaries

This is an approach that can be used with specialized corpora, as specified both in the title to the chapter and in the introduction. Specialized here refers to the domain, i.e. the thematic or topical content of the texts that make up the corpus, not to their genre or register. This point must be stressed, as the method used hinges on there being a clear distinction between the subject matter of the discourse and the remaining, unrelated lexis. Metaphor rests on there being incongruity between the topic/target domain and the vehicle/source domain. The incongruity arises because a word that does not belong to the subject matter being discussed is used when discussing that subject matter.

The procedures for identifying metaphors in corpora described in Section 3 were all adopted for research based on genre-specific corpora. In such a data set, there is no clear-cut division between the subject matter of the data and incongruous subject matter (which may turn out to be metaphorical), as a wide variety of discourse topics

are featured. This means that a word or conceptual domain may be used literally in some texts, and metaphorically in others. It is unlikely that such polysemy would occur within a single text, however, unless special effects such as humour, irony, or cliché were deliberately being sought (Hoey 2005: 82). In a genre-specific corpus, the heterogeneity of discourse topics makes the automatic identification of metaphor extremely problematic. In contrast, the broadly monothematic nature of a specialized corpus makes identifying its core subject matter quite a simple procedure. In this case, it is possible to extend Hoey's (2005: 82) claim that senses of polysemous words tend to avoid each other's textual environments to a discourse setting: it is extremely unlikely that words belonging to the core subject matter of a specialized discourse should be used both literally and metaphorically within that same discourse.[4]

Starting from this premise, then, the topical content of the specialized corpus is treated as a generic *potential* metaphorical topic (for linguistic metaphors) or target domain (for conceptual metaphors found in the corpus), and those words that are unrelated to the discourse topic can be treated as *potentially* metaphorical. Single, "one-shot" occurrences (Lakoff 1987) are *potential* metaphor vehicles, while if semantic or lexical sets can be identified from among the incongruous words, the resulting sets can be said to be *potential* metaphorical source domains. The shift from being *potentially* metaphorical to being confirmed as metaphorical occurs as a result of further investigation of the individual instance in context, which is done by calling up concordances of the relevant word form or lemma (see Section 4.4).

4.3 Establishing the thematic content of the specialized corpus

The data used in this study was downloaded from the Italian government homepage (www.governo.it) over a twelve-month period (June 2006–May 2007), and was to be used for a study of Italian women politicians' metaphorical language (Philip 2009a). The data was stored as five separate corpora (corresponding to the Ministries of Family Policy, Equal Opportunities, Finance, Regional Policy, and Youth Policy and Sport), and within each corpus the different text types – transcribed speeches, press releases and communiqués, and published interviews – were separated into distinct subcorpora. Over a year, some Ministries produce more written output than others, reflecting their relative prominence, with the result that the sizes of the corpora, and the text types included in them, differ considerably. Details of the composition of the subcorpora can be seen in Table1.

In a specialized corpus, one would expect the thematic content to be fairly evident. However, in the case of the corpora discussed here, more than one specialized area may be present, both as a result of the Ministerial remit, and due to the political and

4. There are always exceptions to rules: when teaching a metaphor module in an academic writing in English class, I made use of a text dealing with software design for an architecture application, where the same architecture terms were used (conventionally) to describe how the software was *constructed*, as well as for actual architectural features.

Table 1. Corpus size (running words) and composition

	Subcorpora			
Corpus	**Speeches**	**Communiqués**	**Interviews**	**Total**
Family Policy	32,067	13,658	73,360	119,085
Equal Opportunities	3,107	42,157	–	45,264
Finance	78,926	31,132	–	110,058
Regional Policy	5,172	9,101	–	14,273
Youth Policy and Sport	4,664	30,543	63,121	98,328
	123,936	126,591	136,481	387,008

social climate of the period when the data was collected. When the Ministerial remit is varied, the topics it covers may be quite closely related (as is the case for the Ministry of Finance, which covers trade and commerce, economics, and financial policy), but equally the topics may be quite distinct (there may be some overlaps between the domains of sport and young people's interests, but they are essentially separate). As far as transient socio-political issues are concerned, the Church and religious faith feature prominently in the data for Family Policy, reflecting the conflict between Church and State caused by the civil partnership legislation being discussed when the data was collected (see Table 2). Given this state of affairs, the researcher cannot presume to know *a priori* what lexis and subject matter feature most prominently in the data. It is therefore necessary to find out what words are used, which can be done computationally. In the present study WordSmith Tools version 4 (Scott 2004) was used, although most PC concordance packages on the market provide comparable functionalities.

Content is determined by frequency: the higher the frequency of a word (or lemma), the more central it is to the content of the corpus. Frequency can be calculated as a raw figure (the actual number of occurrences), or as a statistically relevant figure, calculated with reference to a baseline measure. The first stage for either measure is to generate a word-frequency list. This is a very simple procedure and can be done with or without concordance software.[5] Once the list has been generated and displayed by frequency, a cut-off point can be decided and any words occurring below that threshold discarded. The remaining list of frequent words can then be sorted alphabetically – for the sake of convenience – and subsequently grouped into semantically-related sets if finer-grained target domains are sought at this stage.[6]

5. An Internet search for "word frequency generator" or "word frequency list" should lead the researcher to a range of the many freeware applications which can generate such lists, both alphabetically and by frequency.

6. Such groupings can be left until later stages of analysis, as it is the combination of target and source domain groupings which will provide evidence of conceptual metaphors.

The most obvious finding of a raw frequency count is that function words (*the*, *a*, *he*, *of*) appear most often, with content words occurring lower down the list. For an analysis of thematic content, the structure words are of limited interest, as metaphors require content. Structure words can be eliminated manually to leave only content words, but mere frequency of occurrence cannot shed light on the relative importance of those content words to the domain under study. For this reason, it is preferable to determine their frequency relative to other domains by applying a statistical measure that compares the content of the specialized data with less specialized data. In WordSmith tools, this is done through the Key Word function.

Keywords are calculated by comparing the word-frequency list of the corpus being examined with the word-frequency list of a reference corpus. The keyword application computes the frequency in the corpus relative to the number of running words in the same data set, and cross-tabulates the score obtained with that of the frequency of the same word in the reference corpus, relative to the number of running words in the reference corpus (Scott 2004). Words are considered *key* if their occurrence in the researcher's corpus is significantly more frequent than their frequency in the reference corpus, significance being identified by a very low p value (Scott 2000). The default *p* value in WordSmith is 0.000001 (one in 1 million), making classification as key maximally selective, as it is preferable to include fewer, not more words in the keyword list. Table 2 shows the top 20 keywords from the Family Policy corpus, together with frequency information and p value. It should be stressed that there is no priority given to frequency within the keyword calculation: a word either is or is not key, and its position on the keyword list is of minimal importance.

Keyword classifications, being calculated by cross-tabulation of two data sets, are not absolute. Depending on the reference corpus used, different results are obtained. For the purposes of this study, the reference corpus used comprised the combined corpora of Italian political language, not a general reference corpus. The reasons for this choice are given in Section 5.2, together with a comprehensive discussion on the choice of the reference corpus for particular research purposes.

Keywords tell us what the data is about, and provide a good indication of the topics and target domain(s) that will feature in the domain's metaphors. Low-frequency content words (LFCWs), by contrast, are where the metaphor vehicles and source domains will be found. Yet while this is a simple observation, the location of metaphor vehicles/sources is neither as straightforward nor as speedy as the identification of the topics/ target domains. In the first instance, in accordance with Zipf's constant (1935),[7] LFCWs account for a larger proportion of the tokens (running words) in a corpus than

7. On the basis of Zipf's constant, the rank of any word (1 being the most frequent word, 2 the second-most frequent, and so on), when multiplied with its frequency of occurrence (number of tokens), will provide the same figure (the constant), regardless of the rank of the word. In other words, the frequency of any word is inversely proportional to its rank (all words which have the same frequency share the same rank in the frequency table).

Table 2. Top 20 Keywords in Family Policy corpus

		F.P. corpus		Reference corpus			
Keyword*		Frequency	%	Frequency	%	Keyness	P value
1	*famiglia*	951	0.84	1067	0.27	611.51	0.0000000000
2	*più*	650	0.58	878	0.22	315.1	0.0000000000
3	Bindi	441	0.39	457	0.12	313.52	0.0000000000
4	*È*	5005	4.43	13644	3.44	232.77	0.0000000000
5	*non*	1632	1.45	3624	0.91	223.37	0.0000000000
6	*famiglie*	352	0.31	432	0.11	198.55	0.0000000000
7	*perché*	263	0.23	296	0.07	167.82	0.0000000000
8	Rosy	211	0.19	220	0.06	148.7	0.0000000000
9	*figli*	220	0.19	252	0.06	137	0.0000000000
10	*bambini*	183	0.16	199	0.05	122.2	0.0000000000
11	*può*	187	0.17	221	0.06	111.4	0.0000000000
12	*familiari*	132	0.12	145	0.04	86.96	0.0000000000
13	*cattolici*	120	0.11	124	0.03	85.48	0.0000000000
14	*ma*	706	0.63	1649	0.42	77.54	0.0000000000
15	*sarà*	137	0.12	171	0.04	75.23	0.0000000000
16	*chiesa*	99	0.09	103	0.03	69.92	0.0000000000
17	*adozioni*	93	0.08	95	0.02	67.18	0.0000000000
18	*partito*	182	0.16	285	0.07	66.69	0.0000000000
19	*responsabilità*	98	0.09	105	0.03	66.69	0.0000000000
20	*però*	102	0.09	117	0.03	63.36	0.0000000000

* Translations are as follows: (1) family; (2) more; (3) Bindi [Minister's surname]; (4) is; (5) not; (6) families; (7) because; (8) Rosy [Minister's first name]; (9) children [offspring]; (10) children [infants]; (11) can; (12) family members; (13) Catholics; (14) but; (15) will be; (16) church; (17) adoptions; (18) [political] party; (19) responsibility/-ies; (20) though.

do high-frequency words, even though they constitute a small proportion of the word forms (types) present. The dramatic fall-away is illustrated in Figure 1, which shows the distribution of tokens for the top 100 types in the five corpora combined (the most frequent occurs 14,655 times, the hundredth-most frequent, only 393 times). Over ten thousand hapax legomena (i.e. words occurring only once) occur, accounting for 43.27% of the types (10,027 out of a total number of 23,164 types), although they amount to only 2.57% of the total number of tokens in the combined corpora.

Compounding the seemingly interminable number of low-frequency types is the problem that not all low-frequency words are content words, so structure words have to be filtered out. This has to be done manually, unless the researcher is fortunate enough to have access to tools that do the task automatically. Finally, a LFCW is not necessarily metaphorical, meaning that a considerable amount of manual processing has to be done

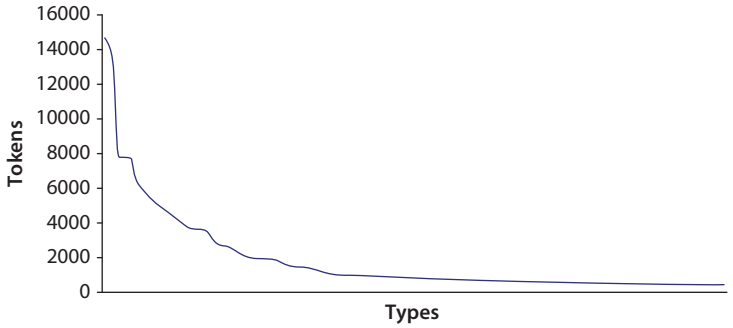

Figure 1. Distribution of tokens in top 100 types (combined corpora)

to find the metaphors. In spite of these problems, though, applying a statistical measure to separate out metaphorical topics/target domains from vehicles/ source domains guarantees the discovery of metaphors that are missed when carrying out partial analysis, ensuring that the analysis covers all metaphors, not just the most prominent ones.

4.4 Grouping and classifying low frequency content words

Once the keywords have been identified (and grouped together, if appropriate), the attention shifts to the LFCWs. Within the LFCWs there will be lexis that is congruous with the keywords, and other lexis that is incongruous. The congruous lexis should be grouped together with the keywords, as it represents alternative wordings referring to the same domain. The incongruous lexis then has to be sorted and grouped by lemma, then by semantic or lexical set (or both). This takes less time if the researcher has access to a lemmatizer to pull together inflected forms, and less still if some form of semantic tagger or classifier can be used; but these are not necessarily available, and were not used in this study.

The most straightforward way to approach the grouping task is to sort the word list alphabetically, which brings related forms together. For minimally inflecting languages such as English, pulling inflected forms together under their respective lemmas is a fairly quick and painless procedure, even without the aid of a lemmatizer. The situation is not quite so straightforward for other languages, however. Italian, while not the most complicated of inflecting languages, presents several complications: nouns, adjectives, participles and most deictics inflect for gender (m./f.) and number (sing./pl.); verbs inflect for six persons in seven tenses, and prepositions merge with the definite article to form *preposizioni articolate* (so *di* + *il* becomes *del*; *in* + *la* becomes *nella*, and so on). These complications make manual lemmatising time-consuming, and it can be tempting to lump all the inflected forms together and work at the more abstract level of the lemma. Should the researcher decide to lemmatize and thus facilitate the arrangement of LFCWs

into semantically related groups, this should be done in such a way that the word forms and their frequencies can still be accessed. As each distinct form collocates differently, the actual forms found in the data should be stored in a spreadsheet or similar database, so that they can be called up in the concordancing software at later stages of analysis.

Once forms have been lemmatized – or even during the procedure – semantic groups will start to coalesce. The more broadly-defined these are the better, as a word that may appear to belong to one semantic class may in fact belong to another, or be potentially a member of both. A selection of the groupings found for the Finance corpus is presented in Table 3. The war grouping – by far the most dominant – was further subdivided into battle, defence and invasion, and victory (see Philip 2010).

Of course, alongside the semantic groups identified, there will be terms that do not seem to fit anywhere in particular. While these are still potentially metaphorical (metaphor vehicle terms), they are unlikely to belong to a conceptual metaphor or metaphor theme. By concordancing these terms, it can be verified whether or not they are in fact metaphors, and if so, whether to consider them conventional or innovative, and to comment on them on the basis of this assessment. The identification of groupings, on the other hand, is potentially indicative of conceptual metaphor at work.[8] The potential metaphors must then be concordanced and are confirmed as metaphorical on the basis of their use. In the data studied, it was found that the same word form or lemma can appear literally in some contexts, and metaphorically in others, so care must be taken in an analysis of metaphors to ensure that any literal uses are kept separate from the non-literal ones.

Having identified the metaphor vehicles (i.e. the words used) and source domains (i.e. the semantic fields the words belong to) in the corpus, their function can then be investigated using the corpus to call up concordances or extended context if required.

Table 3. Low frequency content word groupings in Finance

Semantic field	Examples
Birth	*embrionale* ('embryonic'), *gestazione* ('gestation'), *nascita* ('birth')
Body parts	*cervelli* ('brains'), *ombelico* ('belly button'), *labbra* ('lips')
Death	*soffocamento* ('suffocation'), *strozzature* ('strangulation'), *sterminio* ('extermination')
Emotions	*emotivo* ('emotional'), *sentimenti* ('feelings'), *sensibilizzato* ('sensitized')
Health	*sano* ('healthy') *ferito* ('injured'), *convalescente* ('convalescent')
Hunting	*preda* ('prey'), *caccia* ('hunt')
Risk	*rischio* ('risk'), *sfida* ('challenge'), *salvaguardare* ('to safeguard')
Servitude	*sfruttato* ('exploited'), *servitù* ('servitude'), *sacrificio* ('sacrifice')
War	*battaglia* ('battle'), *conquista* ('conquest'), *sconfiggere* ('to defeat')

8. Editors' note: The procedure described is similar to the "vehicle grouping" described by Cameron et al. (2010: 118–126), which aims to uncover systematicity in metaphor use.

Several lines of investigation open up at this stage, and it is up to the researcher to decide if the corpus-based activity has ceased (except perhaps as a conventional means of locating the examples in the large data set), or if instead the automatic tools can be put to further use. The collocational features of metaphor vehicles can be analysed, for example, as can the co-occurrence of metaphors with keywords (see Philip 2010).

5. Further technicalities

5.1 A note on high and low frequencies

While they are convenient as generic terms, high and low frequency have to be defined clearly in this kind of procedure: a cut-off point must be determined. During the initial, experimental stages of this procedure, raw numerical frequency was used (any term occurring less than five times was deemed "low frequency"), but raw frequency-counts cannot be generalized, and different measures would apply to data sets of different sizes. In refining the procedure, a more robust criterion was established, namely that the cut-off point might correspond to the frequency below which no keywords were extracted. Thus, if the least frequent of the keywords occurs 12 times (as is the case for the European Policy component of the Finance corpus that was used as data for a pilot study), then words occurring 11 times or less are "low frequency" (see Philip 2010). However, this criterion was less successful in the corpora whose content was more diversified, such as Family Policy, whose lowest-frequency keyword occurred 35 times: it is not reasonable to treat any word occurring 34 times or less in a corpus totalling just under 120 thousand words, as "low frequency". There is therefore a middle-cut to consider.

Corpus studies generally look for the presence of collocational patternings forming around a given node, as repetition of patterns is a good indication that the meaning being expressed has become, or is becoming, conventionalized. Having excluded metaphorically motivated terminology from this study, the search for metaphors proper essentially involves searching for the opposite phenomenon, namely *absence* of collocational patternings. An absence of patternings suggests that the node in question has not (yet) gained currency in the discourse with one particular meaning. Strong collocational and phraseological preferences affect the polysemous potential of a word, limiting the likelihood that it will be used both literally and figuratively in the same discourse (see Hoey 2005: 85). As different meanings imply different patterns, the emergence of dominant patternings in a text or discourse makes it less likely that other patterns – and hence, other senses – will occur. When no such dominant patternings can be identified, any of the node's meanings can potentially occur, because the discourse has not expressed a preference for one meaning in particular, and therefore does not block the realization of its other meanings (ibid.). These are favourable circumstances for the realization of metaphorical meanings.

```
un piano strategico di penetrazione commerciale dal 2008 al 2010 ,
azioni di sostegno alla penetrazione commerciale del sistema Italia.
tà di esportazione e di penetrazione commerciale dei nostri imprendi
mento strategico per la penetrazione commerciale delle nostre impres
per la maggior parte di penetrazione commerciale finanziati attraver
azioni più complesse di penetrazione commerciale. Mi auguro che ques
vità di promozione e di penetrazione commerciale. Per l ' anno 2007
nternazionalizzazione e penetrazione commerciale. Il Ministero del C
ossibilità di ulteriore penetrazione commerciale su mercati maturi m
ziativa che rafforzi la penetrazione delle imprese editoriali italia
orte protagonismo nella penetrazione dei mercati esteri. Nella situa
anica strumentale , una penetrazione nel settore dei servizi e in qu
li , di accompagnare la penetrazione sui mercati internazionali con
```

Figure 2. The patterning of 'penetrazione' in the Finance corpus (all occurrences)

In the Italian data studied here, crystallization of collocational patternings was identified in as few as ten concordance lines for the same word form, and occasionally with even fewer. An example is provided in Figure 2, which shows the concordance lines for *penetrazione* (penetration). Here it can be seen that there is a preferred collocate *commerciale* (commercial), as well as an identifiable co-occurrence preference for *penetrazione* with markets and sectors (*mercati*, *settori*); and that these tend to be foreign rather than domestic (*esteri*, *internazionali*).

Those metaphor candidates that occurred in the middle-frequency bands (below the keyword threshold, and above ten occurrences) can be seen to demonstrate stronger co-textual patternings than their lower-frequency counterparts, and as a result begin to consolidate themselves as domain-specific vocabulary or indeed terminology (Philip 2010). Although a more detailed examination of the middle band is beyond the scope of this study, the tendencies observed suggest that further investigation into the crystallization of collocational patternings in specialized discourse may be an interesting and fruitful area for future study.

It must be stressed that the LFCW identification criteria adopted here apply to Italian in particular. The working definition of "low frequency" as corresponding to <10 tokens per type may well vary from language to language, discourse to discourse, and the overall size of the corpora being studied. Languages that are morphologically less complex than Italian, and therefore have few inflected forms, will require the presence of a higher number of tokens per type before patterns begin to crystallize. Every inflected form is delimited by its collocates, but if there are only a handful of inflected forms available – as in English – the number of meanings potentially associated with each is far greater than with a meaning that has scores of inflected forms. This is the main reason why lemmatization is not always advisable: it is known that word forms attract patternings that cannot be generalized to the lemma from which they derive (Sinclair 1996, 1998). The more inflected forms there are of a lemma, the more localized these patterns become, meaning that it is easier to detect them – and the particular meanings they express – when unlemmatized. In order to determine the LFCWs

cut-off point for other languages, therefore, some analysis of middle-frequency terms has to be carried out to verify where crystallization seems to be taking place: the author suggests concordancing one or two of the content words occurring 10, 15, 20, 25 times, as a means of determining that threshold.

5.2 Comparing corpora

Using keywords to identify the subject matter of a corpus is quick and reliable, but only if the two data sets being used (the corpus, and the reference word list) are comparable. In this study, it was decided to use a word-list from the combined political corpora as the reference word-list, necessary for the calculation of keywords in the subcorpora. Some scholars might disagree with this choice, claiming that a word list derived from a large, general reference corpus should have been used instead. This subsection illustrates the different results obtained when the reference data differs.

The prime motivating factor behind using the combined political corpus data, and not a general reference corpus, was to compare like with like. As the *WordSmith Tools* (Scott 2004) help file suggests, "compare apples with pears, or, better still, Coxes with Granny Smiths... and avoid comparing apples with phone boxes!" If detailed and reliable results are to be obtained, it is important to filter out those words that would be key to politics in general, but not key to each ministry's sphere of activity. Different keywords emerge when different reference word lists are used, and the discussion to follow shows precisely why a general reference corpus would not have been suitable for the task in hand.

Although a general reference corpus might seem to be the best choice for a word list for calculating keywords in a smaller corpus, it is its very generality that makes it unsuitable: the general language is simply too unlike a specialized language. Keyness is not just related to subject matter; in fact many stylistic features that might otherwise pass unnoticed can be identified because their statistical significance is revealed in a keyword computation. Of course, comparing the specialized data to a general corpus reveals much about the content of the data, but these differences are not as easily identifiable, nor as relevant, as those that appear in a comparison of two more similar data sets. Comparing specialized with general data not only highlights words that are central to the subject matter of the data set, but also those that are more strictly indicative of style, register, and genre. Table 4 shows the top 50 keywords generated by comparing the Family Policy corpus (see Table 2) with the CoLFIS wordlist (Laudanna et al. 1995), which is derived from a corpus of contemporary written Italian (newspapers, magazines and books). As mentioned in Section 3.3, "top 50" does not mean the 50 most significant keywords, but rather should be interpreted as 10% of the total keywords identified, no keyword being inherently "more key" than another. The words are arranged alphabetically to facilitate reading.

It should be immediately obvious just how few content words appear on this list. There are 500 keywords to trawl through, and the vast majority of them are structural.

Table 4. Top 50 keywords in Family Policy corpus, calculated with reference to CoLFIS

a	to	*il*	the (m. s.)
al	to the	*in*	In
alla	to the	*l*	The
alle	at the	*la*	the (f. s.)
anche	also	*lavoro*	Work
Bindi	Bindi	*le*	the (f. pl.)
che	that	*legge*	Law
ci	[clitic]	*ma*	But
come	like/such as	*ministro*	Minister
con	with	*nel*	in the
da	from	*nella*	in the
dei	of the	*non*	Not
del	of the	*o*	Or
dell	of the	*per*	For
della	of the	*perché*	Because
delle	of the	*più*	More
di	of	*politica*	political (f. pl.); policy; politics
e	and	*politiche*	political (f. pl.); policies; politics
essere	to be	*questo*	This
famiglia	family	*Se*	If
famiglie	families	*Si*	one [reflexive pronoun]
gli	[the m. pl.]	*sono*	am/[they] are
governo	government	*Tra*	Between
ha	has	*Un*	a (m.)
i	the (m. pl.)	*Una*	a (f.)

Much of the information provided here is of interest to stylistics scholars, but if the intention is to establish the subject matter of the data, content words are required. These are somewhat thin on the ground, and still too dispersive for the purpose at hand. Compare the 20 most frequent content words in the two keyword calculations (using the combined political corpora, and using CoLFIS), shown in Table 5.

Using the combined political corpora as the general reference, 60 keywords are obtained, 50 of which (83%) are content words. Words related to the world of politics in general, such as *paese* (nation /country), *politica* (politics), and *Italia* (Italy) are largely filtered out as they are common to the specialized corpus and to the combined corpora. Only when these words are used disproportionately more frequently than normal do they register as key: for instance, *Italia* is key in the Economics corpus, because it features in talk on trade, import, export, branding, and so on, in addition to being used to designate the country itself – which is the use that is common to all five corpora.

Table 5. 20 most frequent content words in the combined political corpora and CoLFIS

Combined political corpora		**CoLFIS**	
famiglia	family	*famiglia*	family
Bindi	Bindi	*servizi*	services
famiglie	families	*ministro*	minister
Rosy	Rosy	*Bindi*	Bindi
figli	children	*famiglie*	families
bambini	infants	*governo*	government
familiari	family members	*politica*	politics
cattolici	Catholics	*politiche*	political
chiesa	church	*lavoro*	work
adozioni	adoption	*legge*	law
responsabilita	responsibility	*paese*	country
partito	party	*figli*	children
vita	life	Rosy	Rosy
anziani	elderly	*diritti*	rights
DICO	DICO*	*vita*	life
coppie	couples	*parte*	part
persone	people	*bambini*	infants
familiare	family (attrib.)	*partito*	party
assegni	[welfare] cheques	*finanziaria*	financial
matrimonio	marriage	*donne*	women

*DIchiarazioneCOngiunta: the name given to the Italian civil partnership legislation

The CoLFIS word list, being far more comprehensive (more general in nature, and far larger in size), makes it possible to identify 500 keywords (the default maximum number in WordSmith; this figure can be changed if so desired). Of these, 303 (60%) are content words. This greater number of keywords does give a more detailed insight into the content of the corpus, but much of what is considered key here is in fact key to political discourse as a genre. We can find words used to talk about politics and politicians, including *costituzione* (constitution), *presidente* (president), *Prodi* (the then Prime Minister), *Margherita* (the then centre-left coalition party). The keywords also reveal the constant presence of persuasive language: there is a plethora of modal and quasi-modal expressions, as well as conditional and future tense inflections. Discourse markers and rhetorical devices also feature strongly. Additionally, there is a noticeable presence of lexis belonging to the sphere of problems, struggles and difficulty. All these features are common to all of the political data analysed here, but particularly to the Ministries that deal with social issues.[9]

9. The financial data is qualitatively different from the other four corpora, for a number of reasons. See Philip (2009a) for a comparative study of the five data sets.

This comparison of keyness serves to illustrate the degree to which the reference corpus influences the identification of keywords. For the purposes of this study, political language in general was of limited interest; rather, the aim was to identify which metaphors were used in one government during a fixed period of time, by different Ministers with different Ministerial remits. Had the intention been to identify political metaphors that were not domain-specific, then each subcorpus should have been compared to a larger data set dealing with the same topic from a range of sources (business and other professional practice, academia, journalism, etc.), thus allowing subject-specific keywords to be filtered out and political and persuasive language to be highlighted. The more similar the data sets are, the easier it is to pinpoint the differences in the keywords that are generated, because there will be fewer keywords (Scott 1997 suggests 40 as a manageable number) and they will be more focused. As a final comment on the matter, there is no reason why only one keyword list should be created for any given study: several keyword lists, each based on comparison with a different reference corpus, will certainly be more revealing than one long, undifferentiated keyword list generated from a general reference corpus.

6. Conclusions

The present chapter has outlined a technique for retrieving metaphor candidates from specialized corpora using computational tools that are cheap, user-friendly, and easily available. Building a corpus from electronic texts is a simple procedure (see Sinclair 2005), and being able to partially automate the location of metaphors in a corpus allows the researcher to concentrate more energy and attention on the analysis of the metaphors once found, rather than on trawling the data manually in search of them.

Concordancing metaphors makes it possible to identify regularities in the phraseological patternings that crystallize around the node. It is argued here that regularity of patterning is a sign of conventionality and that – in specialized corpora at least – it may be advisable to make a distinction between metaphors that are truly figurative and those that are terminological or otherwise domain-specific. Keyword extraction makes it possible to identify metaphorically motivated terms, and separate them from other kinds of metaphor. This makes it possible for a researcher who is unfamiliar with the specialized language in question to assess the force of the metaphorical terms encountered *as they would be assessed by users of that specialized discourse*, which reduces the danger of over-interpreting metaphoricity as a result of unfamiliarity.

Some issues have been left unresolved. A precise cut-off point, below which content words can be defined as "low frequency" has not been established, as it is expected to vary from language to language, and possibly also from domain to domain. Additionally, it is difficult to ascertain the status of low frequency words as metaphorical or simply formulaic when they occur in small corpora, because regularity of patterning can only be identified when forms are repeated a minimum number of

times. Yet the method outlined here opens up the automation of data retrieval to researchers who for whatever reason do not have access to more sophisticated data annotation tools. It is one of several possible approaches to locating metaphors in text corpora (see Berber Sardinha, this volume), and the difficulties encountered, rather than being seen as flaws in the method, should be seen as further opportunities for research into metaphor typologies and the phraseological realization of metaphorical meaning in text.

References

Black, Max. 1993. More about metaphor. In A. Ortony, ed., *Metaphor and Thought*, 2nd edition., 19–43. Cambridge: Cambridge University Press.

Cameron, Lynne. 2003. *Metaphor in Educational Discourse*. London: Continuum.

Cameron, Lynne. 2010. Metaphor in physical-and-speech action sequences. In G. Low, Z. Todd, A. Deignan, & L. Cameron, eds., *Researching and Applying Metaphor in the Real World,* 333–355. Amsterdam & Philadelphia: Benjamins.

Cameron, Lynne & Alice Deignan. 2006. The emergence of metaphor in discourse. *Applied Linguistics* 27 (4): 671–690.

Cameron, Lynne, Robert Maslen, & Graham Low. Finding systematicity in metaphor use. In L. Cameron & R. Maslen, eds., *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*, 116–146. London: Equinox.

Charteris-Black, Jonathan. 2004. *Corpus Approaches to Critical Metaphor Analysis.* Basingstoke & New York: Palgrave Macmillan.

Charteris-Black, Jonathan. 2005. *Politicians and Rhetoric: The Persuasive Power of Metaphor*. Basingstoke & New York: Palgrave Macmillan.

CoLFIS (Corpus eLessico di Frequenza dell'Italiano Scritto). Available at: http://alphalinguistica.sns.it/CoLFIS/CoLFIS_Presentazione.htm [Accessed 2008–10–02].

Deignan, Alice. 2005. *Metaphor and Corpus Linguistics*. Amsterdam & Philadelphia: Benjamins.

Deignan, Alice & Liz Potter. 2004. A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics* 36: 1231–1252.

Gibbs, Raymond W., Jr., Darin L. Buchalter, Jessica F. Moise, & William T. Farrar. 1993. Literal meaning and figurative language. *Discourse Processes* 16: 387–403.

Giora, Rachel. 1999. On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics* 31: 919–929.

Giora, Rachel. 2003. *On Our Mind. Salience, Context, and Figurative Language*. Oxford: Oxford University Press.

Goatly, Andrew. 1997. *The Language of Metaphors*. London: Routledge.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language.* London: Routledge.

Koller, Veronika & Elena Semino. 2009. Metaphor, politics and gender: A case study from Germany. In K. Ahrens, ed., *Politics, Gender and Conceptual Metaphors*, 9–35. Basingstoke: Palgrave Macmillan.

Lakoff, George. 1986. The meanings of literal. *Metaphor and Symbolic Activity* 1: 291–286.

Lakoff, George. 1987. The death of dead metaphor. *Metaphor and Symbol* 2: 143–147.

Laudanna, Alessandro, Anna M. Thornton, Giorgina Brown, Cristina Burani, & Lucia Marconi. 1995. Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. In S. Bolasco, L. Lebart, & A. Salem, eds., *III Giornate Internazionali di Analisi Statistica dei Dati Testuali*, vol. I, 103–109. Rome: Cisu.

Low, Graham, Jeannette Littlemore, & Almut Koester. 2008. Metaphor Use in Three UK University Lectures. *Applied Linguistics* 29 (3): 428–455.

Partington, Alan. 2003. *The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House.* London & New York: Routledge.

Philip, Gill. 2009a. "Non una donna in politica, ma una donna politica": Women's political language in an Italian context. In K. Ahrens, ed., *Politics, Gender, and Conceptual Metaphors*, 83–111. Basingstoke & New York: Palgrave Macmillan.

Philip, Gill. 2009b. Why the prosody isn't always present: Insights into the idiom principle. In M. Mahlberg, V. González-Díaz, & C. Smith, eds., *Proceedings of the Corpus Linguistics Conference CL2009*. Liverpool: University of Liverpool. Available at: http://ucrel.lancs.ac.uk/publications/CL2009/ [Accessed 2011–01–31]

Philip, Gill. 2010. Identifying metaphorical keyness in specialised corpora. In M. Bondi & M. Scott, eds., *Keyness in Text,* 185–204. Amsterdam & Philadelphia: Benjamins.

Philip, Gill. 2011. *Colouring Meaning: Collocation and Connotation in Figurative Language.* Amsterdam & Philadelphia: Benjamins.

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

Scott, Mike. 1997. PC analysis of key words – And key key words. *System* 25 (2): 233–245.

Scott, Mike. 2000. Mapping key words to *problem* and *solution*. In M. Scott, & G. Thompson, eds., *Patterns of Text*, 109–127. Amsterdam and Philadelphia: Benjamins.

Scott, Mike. 2004. *WordSmith Tools version 4*. Oxford: Oxford University Press.

Semino, Elena & Veronika Koller. 2009. Metaphor, politics and gender: A case study from Italy. In K. Ahrens, ed., *Politics, Gender and Conceptual Metaphors*, 36–61. Basingstoke: Palgrave Macmillan.

Sinclair, John M. 1996. The search for units of meaning. *Textus* 9 (1): 71–106.

Sinclair, John M. 1998. The lexical item. In E. Weigand, ed., *Contrastive Lexical Semantics*, 1–24. Amsterdam & Philadelphia: Benjamins.

Sinclair, John M. 2005. Appendix to chapter one: How to make a corpus. In M. Wynne, ed. *Developing Linguistic Corpora: A Guide to Good Practice,* 1–16. Oxford: Oxbow Books. Available at: http://ahds.ac.uk/linguistic-corpora/ [Accessed 2008–10–31]

Zipf, George K. 1935. *The Psychobiology of Language.* Boston: Haughton Mifflin.