

Towards Automatic Retrieval of Idioms in French Newspaper Corpora

Liesbeth Degand and Yves Bestgen

Université catholique de Louvain, Louvain-la-Neuve, Belgium

Abstract

The goal of this paper is to present a procedure for the automatic retrieval of idiomatic expressions from large text corpora. The procedure combines text segmentation techniques and Latent Semantic Analysis. Three indices were computed on the basis of the three-fold hypothesis that: (1) idiomatic expressions should have few neighbours; (2) idiomatic expressions should demonstrate low semantic proximity between the words composing them; (3) idiomatic expressions should demonstrate low semantic proximity between the expression and the preceding and subsequent segments. The result of this procedure shows that we have not yet reached a fully automatic retrieval of idioms from large corpora, but this first trial has shown that we are on the way. The procedure reduces the amount of data to consider to less than a quarter (23.8 per cent) of the original data, of which one-fifth (20.9 per cent) is idiomatic, and nearly 60 per cent (58.8 per cent) is phraseological in nature. In other words, this procedure drastically improves and facilitates hand-based retrieval. In addition, these first results already permit some linguistic exploitation of the retrieved idioms.

1 Introduction

Although figurative language is very frequent overall in both written and spoken language (Sinclair, 1991; Gibbs, 1994), (corpus-based) research on idioms faces a number of obstacles. There is no doubt about the importance of idioms as a whole in texts (especially in written journalism; Moon 1998a), but if one is looking for a particular verbal idiom (e.g. *bite the dust*, *spill the beans*) in a given (multi-million word) corpus, its relative frequency will seldom exceed one per million words (Moon, 1998b). Also, Nicolas (1995, p. 233) found that ‘contrary to received views, at least 90 per cent of V-NP idioms (...) appear to allow some form of (syntactically) internal modification’ (e.g. *be grist for the mill* becoming *be grist for the linguistic mill*). In addition, idiomatization appears to be one of the principal factors in the evolution of language (Chafe, 1970). In

Correspondence:

Liesbeth Degand,
Université catholique de Louvain,
Place B. Pascal, 1,
B-1348 Louvain-la-Neuve, Belgium.

E-mail:

degand@lige.ucl.ac.be

other words, it is part of a dynamic, moving process. It follows from this that extraction from (large) corpora becomes very difficult, if not impossible, because no complete a priori list of idioms, let alone of figurative language, can be established. Hence, if one wants to study the use or frequency of idioms intra-linguistically, inter-linguistically or in any text genre, new techniques need to be developed to retrieve these idiomatic expressions automatically. In this paper, we want to make a proposal towards this goal.

Our aim is to provide a procedure that would enable us to retrieve automatically idiomatic expressions from large text corpora. The core idea of the procedure is that, from a formal point of view, idioms are multiword expressions (Fernando, 1996, p. 3). The aim would thus be to extract all recurring multiword sequences from a large corpus and to eliminate those that are clearly not idiomatic. The procedure combines text segmentation techniques and Latent Semantic Analysis (Landauer *et al.*, 1998). To test the retrieval technique, we will concentrate first on opaque idioms only, i.e. *stable*, *holistic* combinations of two or *more* words whose meaning cannot be derived from its component parts (Weinreich, 1969; Fraser, 1970; Makkai, 1972; Cowie and Mackin, 1975; Cowie *et al.*, 1983; Čermák, 2001). *Stable* here means that the sequences must be composed of strictly co-occurring words;¹ *holistic* means that these linguistic structures are semantically non-compositional (although sometimes analysable, transparent or motivated), involving some form of internal semantic anomaly (Čermák, 2001); and for a start we will concentrate on sequences of four words only (see below). Our idea is to weaken the formal conditions on the sequences, once the procedure has proven its efficiency.

2 Materials

As a basis for our analysis we used a Belgian francophone newspaper corpus of approximately 60 million words² in which we selected all content-articles (leaving out weather forecasts, television programmes, stock exchange reports, etc.). This data set was cut into sentence length segments and lemmatized. Lemmatization and elimination of all digits, special characters, foreign words, proper nouns, and punctuation brought the total number of words down to approximately 11 million. At that stage, we extracted all (non-homographic) function words,³ which reduced the total number to *c.* 5,500,000 words. Finally, we took into account only word sequences that occurred in minimally fifty-word articles, to avoid poor context or context-less segments. This eventually led to a corpus of about 5,400,000 words.

The first step in the idiom extraction procedure is to prepare the text material in order to extract as many idioms as possible. We started by collecting all repeated segments in the data, i.e. all *n*-word strings ($n > 1$) occurring at least three times in the text. The result of this first step is a list of word sequences containing (partial) idioms, but also other kinds of phraseological constructions such as metaphors, routines, collocations,

- 1 We are aware of the fact that this goes contrary to Nicolas' observation about internal modification, but in the testing phase of the procedure, we wanted to avoid as many disturbing factors as possible.
- 2 We used the *Le Soir* corpus containing all daily issues between 1994 and 1997. The corpus is distributed on CD-ROM.
- 3 This means, for instance, that we left in the personal pronoun *son* (his/her) because it is homographic with the noun *son* (sound).

quasi-idioms, and, of course, free word combinations. Different word sequence extraction trials on our text material resulted in the following lists:

three-word sequences: 15,659;
 four-word sequences: 4,442;
 five-word sequences: 677;
 six-word sequences: 106.

We decided to work with the four-word sequences because we expected the best win–loss proportion in this list with respect to the idiom retrieval: most idioms in French are indeed composed of a verbal or adverbial support which combines with one or two complements (one of which is mostly free) (Gross, 1982). Hence, we can expect that all (lemmatized) two- and three-word idioms should be included in the four-word sequences, and that most idioms counting more than four (lemmatized) words will nevertheless be recognizable in the four-word sequence list. The former was, for instance, the case for the expression *faire long feu* extracted from the four-word sequence ‘avoir faire long feu’, or *faire tache d’huile* extracted from ‘pouvoir faire tache huile’. Illustrations of the latter are the expressions *nul n’est prophète en son propre pays* or *avoir plus d’un tour dans son sac* extracted respectively from the four-word sequences ‘nul est prophète son’ and ‘plus tour son sac’.

3 Automatic Idiom Retrieval Techniques

The next step consists in filtering out the idioms from this collection of four-word segments. According to our working definition mentioned above this means extracting all *stable* and *holistic* four-word expressions.

3.1 Extraction from stable expressions

In defining idioms as being fixed or frozen expressions, we take the (restricted) view that ‘an idiomatic phrase cannot be altered; no other synonymous word can be substituted for any word in the phrase, and the arrangement of the words can rarely be modified’ (McMordie, 1972, p. 6). Combining this view with the neighbourhood effect developed by Coltheart *et al.* (1977) gives rise to our first hypothesis.

Hypothesis 1. A sequence with literal meaning has many neighbours, whereas a figurative one has few.

A neighbour is a possible substitute in a given sequence. If this sequence is a word, its neighbours are words of identical length, with their letters in the same position except for one. The word *set*, for instance, has seventeen (orthographic) neighbours (*sea, see, sew, sex, sit, sat, sot, bet, get, jet, let, met, net, pet, ret, vet, wet*), whereas *limb* has five (*lime, limy, lamb, limp, limn*), and *speed* only two (*spend, steed*). Word length and phonemic combinability are determinant for the number of neighbours of a given term. Analogously, we want to hypothesize that lexical fixedness of a sequence is determinant for the idiomaticity of a

segment. Therefore, the more neighbours a segment has, the less fixed it is, and the less idiomatic it will probably be; the fewer neighbours it has, the more fixed it will be, which should increase its chance to be idiomatic. For example, the (idiomatic) segment *wet behind the ears* occurs five times in the Collins WordbanksOnline English corpus,⁴ and only has five neighbours in this corpus. As shown in (1), the five neighbours are all substitutes of the first word in the segment, namely *wet*. The words *behind*, *the*, and *ears* do not show any variation in this segment:

- (1)
- | | |
|-------------------------------|-----------------|
| <i>wet</i> | behind the ears |
| (a sort of) <i>clip</i> | |
| (in the hairline) <i>or</i> | |
| (feeling more) <i>snowy</i> | |
| (to be) <i>scratched</i> | |
| (skipped money) <i>from</i> . | |

On the other hand, the structurally similar (but non-idiomatic) segment *payable on the day* has more than forty-eight neighbours (more than forty⁵ substitutes for the word *payable*, and eight substitutes for the word *day*), of which a sample is given in (2):

- (2)
- | | |
|--------------------------|--------------------------------|
| <i>payable</i> | on the day |
| <i>depending</i> | |
| (his staff) <i>ran</i> | |
| (much) <i>higher</i> | |
| (16 years of) <i>age</i> | |
| (tickets) <i>bought</i> | |
| payable on the | <i>day</i> |
| | <i>boat</i> |
| | <i>shares</i> |
| | <i>best</i> (building society) |
| | <i>death</i> (or bankruptcy) |
| | <i>following</i> (service). |

Applying the same procedure (but automatized) to our four-word sequences means that we expect more variability within the free combinations (more neighbours), than within the idiomatic expressions (fewer neighbours).

3.2 Extraction from holistic expressions

The holistic character or semantic opacity of idioms is generally well-agreed on in the literature. The meaning of an idiom is not the sum of its constituents; an idiom is very often non-literal (for an overview of the literature, see Fernando (1996)). In Čermák's (2001, p. 6) words, idioms involve some 'semantic anomaly'. Hence, from a purely distributional point of view, these idiomatic expressions will combine internally with 'anomalous' words and they will appear in 'unexpected' contexts. For example, a collocate search in the Collins WordbanksOnline⁶ tells us that

- 4 The Collins WordbanksOnline English corpus is composed of 56 million words of contemporary written and spoken text. It is based on the *Bank of English*, a corpus of more than 450 million words developed by Collins Cobuild. More information can be found at http://titania.cobuild.collins.co.uk/boe_info.html.
- 5 There are probably even more substitutes for the word *payable* in this segment, but the online concordance sampler provided by WordbanksOnline (see <http://titania.cobuild.collins.co.uk/form.html>) has a threshold implemented so that no more than forty concordance lines are displayed. If there are more than forty instances found, a random selection is applied.
- 6 The system considers a word to be a collocate as soon as it occurs within four words to the left or right of the selected keyword, i.e. *dust* in the present case.

a word like *dust* co-occurs most often with content words like *dust* (joint frequency of sixty-eight), *gold* (joint frequency of forty-three), and *cloud* (joint frequency of thirty-eight). These co-occurrences build ‘expected’ contexts for the literal meaning. On the other hand, the words *block*, *hot* or *bite* build an ‘unexpected’ context, which hints towards a figurative use of the segment containing *dust*, as in Queen’s: ‘My man got shot, and the block got hot! Another one bites the dust.’

The view we will take here is that ‘expected’ contexts involve semantic regularity and that ‘unexpected’ contexts are semantically anomalous. To measure this semantic regularity versus anomaly, we propose to construct a screener based on Latent Semantic Analysis (LSA), a statistical technique for extracting a ‘semantic space’ from large text corpora, in which the similarity of the meaning of words, sentences or paragraphs can be determined (Landauer *et al.*, 1998; Bestgen and Cabiaux, 2002).⁷ The point of departure of the analysis is a lexical table (Lebart and Salem, 1992) containing the frequencies of every word in each of the documents included in the text material, a document being a text, a paragraph, or even a sentence. For a large corpus segmented in many different documents, this results in a huge but very sparse matrix, which, in our case, consisted of 25,527 terms in 20,054 documents. To build the semantic space, i.e. extract usable semantic dimensions, this frequency table undergoes a singular value decomposition, a statistical technique that can be compared to a factor analysis extracting the most important orthogonal dimensions.⁸ This semantic space was as it were implicitly underlying the data, that is, *latent*, which gave the method its name. Every word has a certain weight on each of the semantic dimensions. Together these weights thus form a vector representing the meaning of the word in the semantic space (see Manning and Schütze, 1999, pp. 558–564, for an illustrative example). In our case, the singular value decomposition was realized with the program SVDPACK (Berry, 1992), and the 250 first eigenvectors were retained. All original words and segments can then be placed into this semantic space. This makes it possible to measure the semantic proximity between each of them. How is this done? As stated above, the meaning of every word is represented by a vector. To calculate the semantic proximity between two words, the cosine between the two vectors that represent them is calculated. The more two words are semantically similar, the more their vectors point in the same direction, and consequently, the closer their cosine will be to unity (which corresponds to coinciding vectors). A cosine of zero shows an absence of similarity, as the corresponding vectors point in orthogonal directions (see Fig. 1). The same procedure can be used to calculate the semantic similarity between two or more given segments.

We propose to use the capacity of LSA to measure the semantic proximity between words and text segments to screen the list of four-word segments from the first step described above. Our hypotheses are as follows.

Hypothesis 2. The semantic proximity between the four-word sequence and its preceding and following segments will be high if the sequence has a literal meaning, and low if it is figurative.

7 More information on LSA can be found at <http://LSA.colorado.edu>.

8 Contrary to a classical factor analysis, the extracted dimensions are very numerous (several hundreds) and non-interpretable. Nevertheless, they can be considered analogous to semantic features used to describe word senses (Landauer *et al.*, 1998).

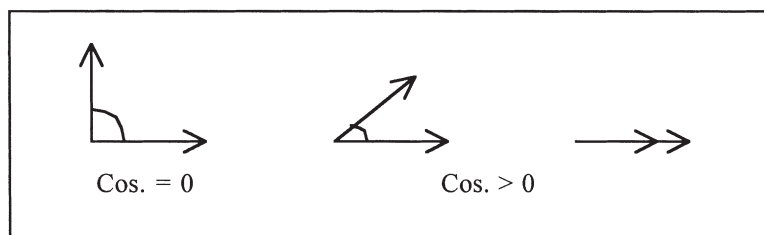


Fig. 1 Vectors showing increasing semantic similarity (cosine between zero and unity).

Hypothesis 3. The semantic proximity between the four words composing the sequence will be high if it has literal meaning, and low if it is figurative.

4 Are Idiomatic and Free Word Combinations Different?

Before we can test these hypotheses on our raw list of four-word sequences, we first have to check whether the implementation of the three above-mentioned hypotheses actually *can* discriminate between idiomatic and free word combinations. To this end, we took a random sample of 60 per cent of the quadruplets, i.e. 2,673 sequences, of which 150 idioms were extracted by hand. The idiomatic nature of the extracted segments was judged by two native speakers, and in cases of doubt they were confronted with an existing list of idiomatic expressions (Rey and Chantreau, 1979). This resulted in two samples: one with 150 idiomatic expressions, and one with 2,523 free word combinations.

For each of the hypotheses an index was computed on the two samples, as follows.

- *N_n*: the number of neighbours in the corpus overall, taking into account the frequency of the neighbours. If the number of neighbours is high, then we should have a free combination; if it is low, we should have an idiomatic sequence.
- MAXCOS: the maximal cosine calculated between each of the four words composing the repeated sequences. A high cosine (close to unity) reveals semantic proximity within the four-word sequence; a low cosine semantic distance.
- COSMAXS1S3: the maximum of two cosines; the cosine between the sequence and S1, composed of the two preceding sentences in the same document, and the cosine between the sequence and S3, composed of the two following sentences within the same document. If the cosine is high, there is high semantic proximity with the preceding or following context (free combination); if the cosine is low, there is low semantic proximity (idiomatic expression).

Table 1 displays the results of the computations for each of the indices. It should be noted that a number of data are missing for the calculation

Table 1 Computation of three indices on idiomatic and non-idiomatic sample

Index	Expression	<i>N</i>	Mean	Std	DF	<i>t</i>	Probability
<i>Nn</i>	Non-idiomatic	2523	40.43	68.74	2671	5.27	0.0001
	Idiomatic	150	10.83	10.38			
MAXCOS	Non-idiomatic	2045	0.53	0.17	2190	8.31	0.0001
	Idiomatic	147	0.41	0.09			
COSMAXS1S3	Non-idiomatic	1937	0.25	0.14	2076	7.20	0.0001
	Idiomatic	141	0.17	0.05			

Std, standard deviation; DF, degree of freedom.

of the MAXCOS and COSMAXS1S3 indices (*N* values). The reason for these missing data is different for the two indices. Before calculating MAXCOS, all highly frequent words within the recurring segments, such as *être* (to be), *faire* (to do), *avoir* (to have), *pas* (not), etc., have been (provisionally) eliminated because they co-occur with nearly every other word in the corpus. Leaving these highly correlating words in the data would artificially boost the cosine towards unity (high semantic proximity). Because of the elimination of these high-frequency words, a number of four-word sequences have shrunk to a single word on which a cosine could not be computed, or have simply disappeared. This process is illustrated in (3) (constructed example):

(3)

<i>Literal segment</i>	<i>Lemmatized segment</i>	<i>Segment without highly frequent words</i>
Je n'ai pas fait la vaisselle.	AVOIR PAS FAIRE VAISSELLE	xxxx xxxx xxxx VAISSELLE
(I didn't do the dishes)	(HAVE NOT DO DISHES)	(xxxx xxxx xxxx DISHES)
Je n'ai pas été faire ...	AVOIR PAS ETRE FAIRE	xxxx xxxx xxxx xxxx
I didn't go doing ...	HAVE NOT BE DO	xxxx xxxx xxxx xxxxx

The reason for the COSMAXS1S3 index to have missing data is that its computation requires the two preceding and the two subsequent segments to be in the same document (in our case, newspaper article). This was not the case for 586 non-idiomatic expressions and nine idiomatic expressions.

A *t*-test was used to test the difference between the means of the non-idiomatic and the idiomatic segments for each index. A look at Table 1 very clearly shows that the two samples significantly differ on all three indices. This means that the three hypotheses are borne out and can be used as parameters to distinguish idiomatic from non-idiomatic expressions.

The question that arises now is whether the technique is useful to extract the idiomatic expressions automatically from the full sample. In other words, if we put all four-word sequences in the same bag again, are we able to extract the idiomatic sequences, and only those?

5 Can We Discriminate Idioms from Non-Idiomatic Sequences?

The results of a univariate discriminant analysis of the sample are displayed in Table 2. It shows that each of the indices can discriminate idiomatic from non-idiomatic sequences, but only partially. In particular, the *Nn* index (number of neighbours) works well to retrieve the idiomatic expressions, as 94 per cent are correctly discriminated. However, the non-idiomatic expressions blur the picture, as only 38.9 per cent are classified correctly. The MAXCOS index (highest cosine between words composing the segment) leads to a loss of 21.9 per cent of the idiomatic expressions (classified as non-idiomatic), and to a 57.7 per cent correct discrimination of the non-idiomatic sequences. Finally, the COSMAXSIS2 index (highest cosine between segment and either two preceding or two following sentences) results in a similar categorization to the COSMAX index: 83 per cent of the idiomatic segments and 51.7 per cent of the non-idiomatic ones are correctly classified.

We then performed a multivariate discriminant analysis using all three indices and this clearly improved the overall result. Table 3 shows that 95.7 per cent of the idiomatic sequences and 68.2 per cent of the non-idiomatic ones are discriminated correctly.

In other words, before the multivariate discriminant analysis the sample examined consisted of 2,673 sequences of which 5.6 per cent are idiomatic. The multivariate discriminant analysis reduces this sample to examine to 637 sequences (504 + 133 segments classified as being idiomatic), of which 20.9 per cent are idiomatic. Instead of encountering one idiom every twenty segments, the rate rises to one idiom every five segments. At this stage, the price to pay is a loss of 4 per cent of the idiomatic expressions (six expressions), which were wrongly classified as being non-idiomatic and a loss of 7.3 per cent (eleven expressions) owing to missing data.

Table 2 Univariate discriminant analysis of three indices

	Classified as idiomatic	Classified as non-idiomatic	Total
<i>Nn</i>			
Idiomatic	141 (94%)	9	150
Non-idiomatic	1542	981 (38.9%)	2523
<i>MAXCOS</i>			
Idiomatic	116 (78.9%)	31	147
Non-idiomatic	865	1180 (57.7%)	2045
<i>COSMAXSIS3</i>			
Idiomatic	117 (83%)	24	141
Non-idiomatic	936	1001 (51.7%)	1937

Table 3 Multivariate discriminant analysis of all three indices

<i>Nn/Maxcos/CosMax13</i>	Classified as idiomatic	Classified as non-idiomatic	Total
Idiom	133 (95.7%)	6	139
Non-idiom	504	1082 (68.2%)	1586

Table 4 Type of linguistic structures classified as being idiomatic ($N = 504$)

Linguistic structure	Frequency	Example
Free word combination	39.3% (198)	accepter de vivre un peu moins (accept to live a little less); l'album n'est pas sorti (the album didn't come out)
Collocation	21% (106)	[après] une année de bons et loyaux services ([after] a year of good and faithful service); l'attentat coûta la vie au président (the attempt cost the president's life)
Title, proper noun phrase	17.5% (88)	l'Association générale des journalistes professionnels (the general association of professional journalists); la Bibliothèque royale au boulevard de l'empereur (the Royal Library at the boulevard de l'empereur)
Expression/chunk	11.7% (59)	les choses ne tournent pas rond (things do not turn round/things do not run as they should), le coup d'envoi sera donné [à] (kick-off will be given [at]); aller au bout de ses idées (to go until the end of one's ideas)
Metaphor	5.2% (26)	avoir des mots très durs (to have hard words), avoir été la proie des flammes (have been the victim of the flames)

9 For ease of illustration, plausible strings have been reconstructed from the lemmatized sequences. For example, *accepter vivre peu moins* (accept live little less) has been reconstructed as *accepter de vivre un peu moins* (accept to live a little less); *année bon loyal services* (year good faithful services) has become *[après] une année de bons et loyaux services* ([after] a year of good and faithful service).

10 Collocations that are clearly metaphorical in nature have been counted as metaphors rather than collocations.

11 This category comprises a list of expressions that were not counted as idiomatic by Rey and Chantreau (1979) but that constitute nevertheless a recurrent, holistic string.

12 Of course, this does not mean that the procedure is able to retrieve all phraseological constructions from the data. Other, additional constraints should apply on the retrieval of collocations, for instance, but this lies beyond the scope of this paper.

To try to improve this result, we examined the list of 504 four-word sequences that were (wrongly) classified as idiomatic. Table 4 gives the distribution of the various types of linguistic structures encountered, along with a number of illustrative examples.⁹

Interestingly, only 39.3 per cent of the segments are indeed free word combinations that were wrongly classified as being idiomatic, but approximately the same number (37.9 per cent) are actually expressions that are rather phraseological in nature, namely, collocations,¹⁰ chunks,¹¹ and metaphorical expressions. As these structures share a number of features with idiomatic expressions (a loose degree of fixedness and semantic opacity), it is not illogical that they are retrieved too.¹² The remaining 17.5 per cent are segments referring to the name or title of associations, books, films, institutions, etc.

6 Conclusion and Future Research

In this paper, we have developed a procedure aimed at retrieving automatically idiomatic expressions from large text corpora. The procedure is a combination of text segmentation techniques and Latent Semantic Analysis. Three indices were computed on the basis of the three-fold hypothesis that: (1) idiomatic expressions should have few neighbours; (2) idiomatic expressions should demonstrate low semantic proximity between the words composing them; (3) idiomatic expressions should demonstrate low semantic proximity between the expression and the preceding and subsequent segments. The result of this procedure shows that we have not yet reached a fully automatic retrieval of idioms from large corpora, but this first trial has shown that we are on the way. As a

matter of fact, the procedure reduces the amount of data to consider to less than a quarter (23.8 per cent) of the original data, of which one-fifth (20.9 per cent) is idiomatic, and nearly 60 per cent (58.8 per cent) is phraseological in nature. In other words, this procedure drastically improves and facilitates hand-based retrieval.

In addition, these first results already permit some linguistic exploitation of the retrieved idioms. One of the often heard complaints with respect to the corpus-based studies of idioms is that the individual occurrence of idioms is too low to authorize any conclusions on their contextualized use (Moon, 1998b). In this context it is noteworthy that several idioms retrieved from our newspaper corpus occur more than ten times (*donner ses lettres de noblesse* (eleven), *chercher midi à quatorze heures* (eleven), *voir plus loin que le bout de son nez* (twelve), *prendre son destin en main* (twelve), *mettre le feu aux poudres* (fourteen), *ne pas y aller par quatre chemins* (fifteen)), with outliers close to twenty (*donner le feu vert* (seventeen), *être réduit à sa plus simple expression* (seventeen)), and above thirty (*ne pas dire son dernier mot* (thirty-six)). These frequencies should enable us to give reliable descriptions of the context of use of these expressions. Thus, the procedure reveals itself as a truly corpus-driven method, disclosing material that could not be studied otherwise.

Finally, in the same vein, the procedure can already be used as a tool that could enable us to measure the level of idiomaticity of a given language and/or text genre, not only of the newspaper corpora under investigation, but also of other pieces of discourse, because the calculated semantic space is also exportable to other corpora. This is of particular interest to contrastive phraseology. Applying this technique to parallel corpora could give insight into the translatability of idioms, as well as give constraints on the contexts in which 'equivalent' idioms can occur in different languages. The transferability of the technique to other languages than French is, however, the topic of future research.

Acknowledgements

The two authors are research fellows of the Belgian National Fund for Scientific Research (FNRS). This research was supported by grant FRFC 2.4535.02 from the Belgian Federal Government.

References

- Berry, M. W. (1992). Large scale singular value computation. *International Journal of Supercomputer Application*, 6: 13–49.
- Bestgen, Y. and Cabiaux, A. F. (2002). L'analyse sémantique latente et l'identification des métaphores. *Actes de la 9ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles*. Nancy: INRIA, pp. 331–7.
- Čermák, F. (2001). Substance of idioms: perennial problems, lack of data or theory? *International Journal of Lexicography*, 14(1): 1–20.
- Chafe, W. (1970). *Meaning and the Structure of Language*. Chicago, IL: University of Chicago Press.

- Coltheart, M., Davelaar, E., Jonasson, J., and Bessner, D. (1977). Access to the internal lexicon. In: Dornic, S. (ed.), *Attention and Performance VI*. New York: Academic Press, pp. 535–55.
- Cowie, A. P. and Mackin, R. (1975). *Oxford Dictionary of Current Idiomatic English. 1, Verbs with Prepositions and Particles*. Oxford: Oxford University Press.
- Cowie, A. P., MacCaig, I. R., and Mackin, R. (1983). *Oxford Dictionary of Current Idiomatic English. 2: Phrase, Clause and Sentence Idioms*. Oxford: Oxford University Press.
- Fernando, C. (1996). *Idioms and Idiomaticity*. Oxford: Oxford University Press.
- Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, 6: 22–42.
- Gibbs, R. Jr (1994). Figurative thought and figurative language. In: Gernsbacher, M. A. (ed.), *Handbook of Psycholinguistics*. New York: Academic Press, pp. 411–46.
- Gross, M. (1982). Une classification des phrases ‘figées’ du français. *Revue Québécoise de Linguistique*, 11(2): 151–85.
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2–3): 259–84.
- Lebart, L. and Salem, A. (1992). *Statistique Textuelle*. Paris: Dunod.
- Makkai, A. (1972). *Idiom Structure in English*. The Hague: Mouton.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- McMordie, W. (1972). *English Idioms and How to Use Them*. Oxford: Oxford University Press.
- Moon, R. (1998a). *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon Press.
- Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In: Cowie, A. P. (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 79–100.
- Nicolas, T. (1995). Semantics of idiom modification. In: Everaert, M., van der Linden, E.-J., Schenk, A., and Schreuder, R. (eds), *Idioms: Structural and Psychological Perspectives*. Hillsdale, NJ: Lawrence Erlbaum, pp. 233–52.
- Rey, A. and Chantreau, S. (1979). *Dictionnaire des Expressions et Locutions*. Paris: Le Robert.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Weinreich, U. (1969). Problems in the analysis of idioms. In: Puhvel, J. (ed.), *Substance and Structure of Language*. Berkeley, CA: University of California Press, pp. 23–81.

