# Updating LSP dictionaries with collocational information

Katerina T. Frantzi

Department of Mediterranean Studies,
University of the Aegean
Dimokratias 1, 85100, Rhodes, Greece
frantzi@rhodes.aegean.gr

**Abstract**

Despite the big amount of general language dictionaries in electronic form, those coming from specialised areas are still "under construction". There are two main reasons for this: firstly, the need for these dictionaries was/is less essential than the need for general language dictionaries, since these were/are aiming mainly to specialists, and secondly many specialised areas are changing over time, resulting to dictionaries that need continuing updating. Due to this, techniques that improve the automatic or semi-automatic construction and updating of specialised dictionaries are and will always be welcome.
In this work we are concerned with the updating of dictionaries for Languages for Special Purposes (LSPs) with information coming from collocations. The collocations to be used are extracted from LSP corpora of not necessarily big size.

## 1 Introduction – collocations

The big number of applications for collocations (dictionary construction, translation, language learning, etc.), makes them an interesting area to work on. The availability of corpora in electronic form has given a great deal of help to this kind of research since we are now able to work with real data. English is not any more the only language with electronic corpora, though it owns the greatest deal. Also, although most electronic corpora describe the general language, corpora of languages for special purposes (LSPs) become more and more available.

Firth, (Palmer 1968), introduced the meaning of a collocation when discussing about senses. He suggested that part of the sense of a word depends on its neighbour words in texts: "You shall know a word by the company it keeps", (Palmer 1968:179). This "company" is what he named collocation, and kept it very important for understanding words.

It is quite some time now that linguists have shown interest in collocations (Jones and Sinclair 1974), and various definitions have been given. Some allow collocations to only consist of two words, while others of much more. Some care about what information collocations can give us on semantics, others on syntax or grammar. Some accept common words, others not. Some allow collocations to cross a comma, others not. Regarding interrupted collocations, there are differences as for the size of the gap(s) among the collocates. Despite all the differences, collocations are arbitrary, recurrent and cohesive lexical clusters, and depend on the language (Smadja 1993). We adopt the collocation definition given by Sinclair and Carter agreeing for a collocation to be the occurrence of two or more words within a short space of each other in a text (Sinclair and Carter 1991).

As mentioned above, collocations depend on the language and sublanguage they are found. They actually play an important role in sublanguages (Frawley 1988; Ananiadou and McNaught 1995). The study of collocations in general language needs large corpora since phenomena in general language are sparse: in the Brown Corpus we only have two instances of "cups of coffee", five of "for good" and seven of "as always" (Kjellmer 1994). However, when we talk about LSPs, things are easier as for the size of the corpus which can be a lot smaller since information there is dense.

Early work on collocation extraction was determinant. Choueka et al. were among the first to use frequency of occurrence for recognising collocations (Choueka et al 1983). The work of Nagao and Mori was also

based on frequency of occurrence but they also considered the length of collocations to be extracted, giving priority to longer ones (Nagao and Mori 1994). Church and Hanks were the first to use association ratio (Church and Hanks 1990), a measure based on mutual information first expressed by Fano (Fano 1961). They cared about the semantic relations of the word-pairs they recognised, which could be interrupted by other words. On mutual information is based the work of Kim and Cho (Kim and Cho 1993), which extent it to three words, but in a different way than that originally defined by Fano. Collocation extraction is still an interesting issue for researchers (Kilgarriff and Tugwell 2001; Kim et al. 2001).

Collocations can be divided to those that do not appear as part of other longer collocations and those that they do. The latter we call *nested collocations*. For example, in Computational Linguistics, "Natural Language" is a collocation itself, but is also part of the longer collocation "Natural Language Processing". Three important works that mention the problem of nested collocations are those of Smadja, Kita et al., and Ikehara et al. Xtract, based on frequency of occurrence, recognised as collocations only those expressions of the greatest length (Smadja 1993). It did not extract collocations that were part of others. The work of Ikehara et al., which was based on Nagao and Mori's work, only accepted those that were found with satisfying frequency as not-nested (Ikehara et al. 1995). The problem of nested collocations was a big a concern for Kita et al. These accepted a nested collocation when it also appeared as not-nested with satisfying frequency (Kita et al. 1994).

**2 Updating the dictionary**

We deal with the updating of LSP dictionaries for the Greek language. We use nested collocations to get the information in a way easier than looking directly into the corpus, which can be very time consuming. *C-value* is used for the extraction of collocations from LSP corpora. *C-value* has been initially constructed and used for the extraction of English collocations (Frantzi et al. 2000). It has been also applied to Japanese language (Mima et al. 2001). In this work we will be using it for Greek collocations and the updating of Greek dictionaries.

Let us remind that *C-value* pays particular attention to nested collocations. When applied to the "Artillery Firing Military Rule Book" ("Στρατιωτικός Κανονισμός Πυροβολαρχίας Βολής", the corpus we will be using) one of the collocations it extracts is the "*Διορθώσεις ως προς τη γραμμή βολής*". So it does to "*γραμμή βολής*" which is a nested to the previous one collocation, but also stands as a collocation by itself. We need such a method since we will use nested collocations to get the information for updating the dictionary.

When *C-value* is applied to an expression, it considers the following parameters:
1. The length of the expression (in terms of number of words). The longer the expression, the more important.
2. The frequency of occurrence of the expression in the corpus. The bigger the frequency the more important the expression.
3. Whether the expression appears as nested, and if yes the number of the different longer collocations that contain it. The number of times it is found in these longer collocations is also a considered parameter.

Let us remind that *C-value* is evaluated as follows (*a* is the expression we examine):

1. *C-(value)(a)=0*

if the expression is part of one longer collocation and its frequency of occurrence is the same as this longer collocation's frequency. In this case the examined expression is not a collocation by itself.

2. *C-value(a)= (|a| - 1)n(a)*

if the expression is not part of any longer collocations.
$|a|$ is the size of the expression *a* in terms of number of words,

*n(a)* is the frequency of occurrence of the expression *a* in the corpus.

3. *C-value(a)= (|a| - 1)(n (a) – t(a)/c(a))*

if the expression is part of longer (more than one) collocations.
*c(a)* is the number of these longer collocations that include the expression *a*,
*t(a)* is the total frequency of the expression *a* as part of these longer collocations.

After extracting the collocations we group them and choose a *group* to start with. Attention should be taken when grouping the collocations. If for example we only group them alphabetically based on the first word, then we could miss out members of the group and as a result possibly useful information. Which group of collocations to start with is up to the application. A group of collocations that would be used to update the dictionary could be the following:

*παράγγελμα βολής*
*αρχικό παράγγελμα βολής*
*αρχικό παράγγελμα βολής μοίρας*
*αρχικό παράγγελμα βολής πυροβολαρχίας*
*αρχικό παράγγελμα άμεσης βολής*
*έντυπο καταγραφής παραγγέλματος βολής*
*εκφώνηση αρχικού παραγγέλματος βολής*
*αρχικό παράγγελμα βολής δεξιού ουλαμού*
*αρχικό παράγγελμα βολής αριστερού ουλαμού*
*αρχικό παράγγελμα βολής κεντρικού ουλαμού*
*αρχικό παράγγελμα άμεσης βολής πυροβολαρχίας*
*αρχικό παράγγελμα άμεσης βολής δεξιού ουλαμού*
*αρχικό παράγγελμα άμεσης βολής αριστερού ουλαμού*
*αρχικό παράγγελμα άμεσης βολής κεντρικού ουλαμού*

The algorithm for updating the dictionary is the following:

```
L: existing LSP dictionary;
entry_L(.): an entry in L;
Extract collocations from the LSP corpus using C-value;
Group collocations creating collocation_groups;
for each collocation_group cg from collocation_groups
        for each collocation c in cg
                length(c)= number of words of c
        max_length = max(length(c)) where collocation c in cg
        new_c = collocation c in cg with length(c)=min(length(.))
        if entry_L(new_c) = 0
                create entry_L(new_c)
        info_length = length(new_c)
        while info_length < max_length + 1
                for each collocation c from cg with info_length = length(new_c)+1
                        check c for new information
                        update entry_L(new_c)
                end_for
```

*info_length = info_length +1*

end_while

end_for

The choice of *C-value* as the method for extracting the collocations is critical since it deals with nested collocations, the type of collocations we need for getting the information. Let us now assume the following imaginative group of collocations from the collocation list:

*a b*
*a b c*
*a b d*
*a b f e*
*a b c g*
*a b f g h*

where *a b c d e f g h* words.

We take the collocation of the smallest length. In our example the "*a b*". If the collocation "*a b*" does not yet exist in the lexicon a new entry is created. Now we consider the collocations of length the next smallest, (in terms of number of words). In our case the "*a b c*" and the "*a b d*". We can start with "*a b c*" considering the word "*c*" in terms of the information it can give us on grammar, syntax or semantics (depending on the type of dictionary we want to update). We continue with "*a b d*" and the grammatical, syntactical or semantical information that the word "*d*" gives for collocation "*a b*". Then we move to collocations (of the same group always) of the next smallest length, that is the "*a b f e*" and the "*a b c g*". We do the work we did before, so we consider "*f e*" as for the information it can give for the collocation "*a b*". For the collocation "*a b c g*" we consider the fact that "*a b c*" is a nested collocation we have already checked and add the information given by word "*g*", and of course any new information acquired by the word combination "*c g*". We finish with the collocation "*a b f g h*", where we take information from the word combination "*f g h*".
When a collocation group is over we can move to the next collocation group.

The method is semi-automatic since the machine, the domain expert and lexicographer need to cooperate. The human factor is necessary for the evaluation of information coming from the collocation under consideration. It is the domain expert and the lexicographer to judge which information is useful to be used and which not.

**3 Application**

The method is applied to the "Artillery Firing Military Rule Book" ("Στρατιωτικός Κανονισμός Πυροβολαρχίας Βολής") of about 35,000 words. Since we are working with an LSP corpus we can use a small corpus. With a general language corpus things would be a lot harder in terms of its size since phenomena in that case are sparse. No tagging has been applied on the corpus. The implementation was done in Linux. Table 1 shows a sample of it.

At first, collocation extraction is taking place using *C-value*. In this application we extract expressions of 2 to 7 words. This is a variable and changes according to application. The extracted collocations are ordered according to their *C-value.* A threshold can be applied to only allow those expressions above a value to be extracted and therefore proceed to the next stage. A threshold could have also been applied to the frequency of occurrence of the candidate expressions.

Table 2 shows a sample of the list with the extracted collocations. The first column gives the *C-value* for the expression shown on fifth column. The fourth column gives the frequency of occurrence of the expression. The third column gives the number of (longer) expressions that contain the current expression while the second the total frequency of the expression in these longer ones. Expressions on Table 2 have

been chosen such that differences between *C-value* and frequency of occurrence can be noticed. We can see for example that long expressions despite their low frequency are valued high by *C-value*, e.g. "*σε διορθώσεις ως προς τη γραμμή βολής*", and "*το γέμισμα είναι ίσο με το βελησνεκές*". Those expressions are domain-dependent, and for that they are (correctly) valued high. On the contrary expressions such "*και στο*" ("*and to*"), "*τη γωνία*" ("*the angle*"), and "*αυτό είναι*" ("*this is*"), are valued more by pure frequency of occurrence.

---

ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑ ΤΟΥ ΠΥΡΟΒΟΛΙΚΟΥ ΜΑΧΗΣ

1. Συνεργασία για την Εκτέλεση Βολής Πυροβολικού.

Το πρόβλημα της υποστηρίξεως δια πυρών μιας Μονάδας ελιγμού επιλύεται με τις συντονισμένες προσπάθειες του παρατηρητή, του Κέντρου Διευθύνσεως Πυρός (ΚΔΠ) και της Πυροβολαρχίας Βολής
(Σχ. 1). Τα τρία αυτά τμήματα του Πυροβολικού, πρέπει να είναι συνδεδεμένα με επαρκές δίκτυο επικοινωνιών. Το ισχύον δόγμα απαιτεί να ενεργούν με ταχύτητα και να καταβάλλουν συνεχώς προσπάθειες μειώσεως του απαιτούμενου χρόνου, για την αποτελεσματική εκτέλεση μιας αποστολής βολής.

α. Παρατηρητής.

Ο παρατηρητής είναι «τα μάτια» του Πυροβολικού Μάχης. Αναζητά και προσδιορίζει τη θέση κατάλληλων για το Πυροβολικό στόχων, μέσα στη ζώνη παρατηρήσεώς του. Για να προσβάλλει ένα στόχο, διαβιβάζει την αίτηση βολής και όταν απαιτείται εκτελεί κανονισμό της βολής. Επιτηρεί τα πυρά του και παρέχει στοιχεία στο ΚΔΠ.

β. Κέντρο Διευθύνσεως Πυρός.

Το ΚΔΠ αποτελεί τον «εγκέφαλο» του Πυροβολικού. Λαμβάνει την αίτηση βολής του παρατηρητή, προσδιορίζει στοιχεία βολής και τα μετατρέπει σε παραγγέλματα βολής, τα οποία διαβιβάζει στα πυροβόλα. Εκτελεί δηλαδή την τεχνική διεύθυνση του πυρός. Λόγω των μεγάλων αποστάσεων μεταξύ των μονάδων πυρός (πυροβολαρχιών) και των απαιτήσεων για την ταχεία παροχή πυρών υποστηρίξεως, η τεχνική διεύθυνση του πυρός διεξάγεται συνήθως στο ΚΔΠ της Πυροβολαρχίας. Το ΚΔΠ Μοίρας παρέχει τακτική διεύθυνση του πυρός (τρόπο προσβολής των στόχων) και παρακολουθεί όλα τα δίκτυα βολής. Επιπλέον βοηθά τα ΚΔΠ των Πυροβολαρχιών στην τεχνική διεύθυνση του πυρός, παρέχοντας σε αυτά στοιχεία βολής για τα σχέδια πυρός και ενεργώντας σαν εφεδρικό ΚΔΠ, όταν απαιτηθεί.

Σχήμα 1. Συνεργασία για την Εκτέλεση Βολής Πυροβολικού.

γ. Πυροβολαρχία Βολής.

---

Table 1 Sample of the corpus.

Table 3 shows how the method behaves with nested expressions. We can see that, if instead of *C-value* we used frequency of occurrence, and in order to give a value to a candidate expression we were subtracting from its frequency the summation of its frequency when part of longer expressions, we would underestimate quite a few important expressions.

The extracted list is expected to contain "useless" expressions, like **"***και στο***"** (**"***and to***"**) or **"***αυτό είναι***"** (**"***this is***"**). However according to Kjellmer no extracted expression can easily -if at all- be characterised "useless" (Kjellmer 1994). His dictionary of English collocations incorporates everything that has been extracted with no characterisation as **"**correct**"** or **"**wrong**"**. However, we could use a part-of-speech tagger to only allow expressions of a particular form. This way we would eliminate some expressions we do not want but could also lose some we do. What we do depends on the application.

| *C-value(a)* | *t(a)* | *c(a)* | *f(a)* | *extracted collocation* |
|---|---|---|---|---|
| 8.42206 | 0 | 0 | 3 | το γέμισμα είναι ίσο με το βελησεκές |
| 8.42206 | 0 | 0 | 3 | σε διορθώσεις ως προς τη γραμμή βολής |
| 8.42206 | 0 | 0 | 3 | ο παρατηρητής προσδιορίζει τη θέση του στόχου |
| 8.42206 | 0 | 0 | 3 | ο παρατηρητής βλέπει τη διάρρηξη του βλήματος |
| 8.42206 | 0 | 0 | 3 | με τα διόπτρα του μετρά τη γωνιακή |
| 8.42206 | 0 | 0 | 3 | M10/M17 η διόρθωση ως προς τη γραμμή |
| 8.42206 | 0 | 0 | 3 | ίσο με το βελησεκές σε χιλιάδες μέτρων |
| 8.42206 | 0 | 0 | 3 | η διόρθωση ως προς τη γραμμή πυροβόλα |
| 8.42206 | 0 | 0 | 3 | η γωνία γνωστού σημείου παρ στόχου είναι |
| 8.42206 | 0 | 0 | 3 | για την εκτέλεση απ' ευθείας δραστικής βολής |
| 8.42206 | 0 | 0 | 3 | γέμισμα είναι ίσο με το βελησεκές σε |
| 8.42206 | 0 | 0 | 3 | από το αβάκιο M10/M17 η διόρθωση ως |
| 8.42206 | 0 | 0 | 3 | αβάκιο M10/M17 η διόρθωση ως προς τη |
| 8.33333 | 8 | 3 | 11 | τη γωνία |
| 8.33333 | 8 | 3 | 11 | και στο |
| 7.92481 | 12 | 2 | 11 | ο παρατηρητής πρέπει |
| 7.66667 | 7 | 3 | 10 | αυτό είναι |
| 7.66667 | 10 | 3 | 11 | στη ζώνη |
| 7.66667 | 10 | 3 | 11 | παρατηρητή να |
| 6.96578 | 0 | 0 | 3 | ο διοικητής της μονάδας ελιγμού |
| 6.96578 | 0 | 0 | 3 | με συσχέτιση προς γνωστό σημείο |
| 6.96578 | 0 | 0 | 3 | για την προσβολή του στόχου |
| 6.96578 | 0 | 0 | 3 | για την εκτέλεση της αποστολής |
| 5.61471 | 0 | 0 | 2 | υπολογισμός αποστάσεως κατά μήκος της γραμμής παρατηρήσεως |
| 5.61471 | 0 | 0 | 2 | τον ακριβή προσδιορισμό της θέσεως των στόχων |

Table 2 Sample of the list with the extracted collocations.

Let us now see an example from our corpus, on how we get the information for updating the dictionary. We have already extracted the collocation list. Assume the collocation group we work with, is the following:

Length: 2
*στοιχεία βολής*
Length: 3
*υπολογισμός στοιχείων βολής*
*καταγραφή στοιχείων βολής*
Length: 4
*στοιχεία βολής προσβολής στόχου*
*μέθοδος υπολογισμού στοιχείων βολής*
Length:5
*στοιχεία βολής από επισήμανση ακριβείας*
Length: 6
*υπολογισμός στοιχείων βολής με χρήση PC32F*
*υπολογισμός στοιχείων βολής με χρήση TI59*
*στοιχεία βολής με χρήση μετεωρολογικών στοιχείων*
*υπολογισμός στοιχείων βολής με χρήση laser*
*υπολογισμός στοιχείων βολής από τον παρατηρητή*
Length: 7
*αναγωγή στοιχείων βολής λόγω χρήσης διαφορετικού πυροσωλήνα*
*εξαγωγή στοιχείων βολής με χρήση αβακίου M17*
Length: 8
*χρήση ΣΕΠ Πυθαγόρας για τον υπολογισμό στοιχείων βολής*

The collocation we are dealing with is the "*στοιχεία βολής*". *Length* is taken in terms of number of words. The collocation is met in the corpus under two forms: "*στοιχεία βολής*" and "*στοιχείων βολής*". We take these two as the same collocation and then we consider the two collocations of *Length=3*. The domain expert and the lexicographer need to evaluate the information taken from each of the two words "*υπολογισμός*" and "*καταγραφή*". The domain expert has to decide whether the expression "*υπολογισμός στοιχείων βολής*" is a collocation or not. If it is, and the information of the word "*υπολογισμός*" regarding the "*στοιχείων βολής*" does not exist in the dictionary then the dictionary has to be updating by the lexicographer with the information given by the domain expert. The same happens for the other collocation of length=3, the "*καταγραφή στοιχείων βολής*". When we finish with collocations of *Length=3*, we move to those of *Length=4*, in our example the "*στοιχεία βολής προσβολής στόχου*" and the "*μέθοδος υπολογισμού στοιχείων βολής*". The first collocation "*στοιχεία βολής προσβολής στόχου*" does not directly relate to any of the two collocations of length=3, and so the "*προσβολής στόχου*" will be treated by the domain expert and the lexicographer as the words "*υπολογισμός*" and "*καταγραφή*" of the previous stage. The collocation "*μέθοδος υπολογισμού στοιχείων βολής*" will have to update the information of the previously checked collocation "*υπολογισμός στοιχείων βολής*", so the latter will be taken under consideration. The method continues with the same simple way until we reach and use the collocation with the greatest length in the group, in our example the "*χρήση ΣΕΠ Πυθαγόρας για τον υπολογισμό στοιχείων βολής*" with *Length*=8.

| *C-value(a)* | *t(a)* | *c(a)* | *f(a)* | *Extracted collocation* |
|---|---|---|---|---|
| 3 | 4 | 2 | 5 | γωνιομετρικό όργανο |
| 8.8 | 11 | 5 | 11 | διόρθωση βελησκούς |
| 42.7 | 33 | 10 | 46 | δραστική βολή |
| 10.2857 | 12 | 7 | 12 | εκρηκτικό βλήμα |
| 10.8 | 11 | 5 | 13 | επισήμανση ακριβείας |
| 11.8872 | 10 | 4 | 0 | ευθεία δραστικής βολής |
| 23 | 26 | 13 | 25 | ζώνη παρατηρήσεως |
| 49.5 | 50 | 20 | 52 | κατά διεύθυνση |
| 8.75 | 10 | 8 | 10 | πρώτη διάρρηξη |
| 18.7143 | 23 | 7 | 22 | πρώτο βλήμα |
| 11.4118 | 10 | 17 | 12 | πυροβόλα στόχου |
| 18.8571 | 15 | 7 | 21 | σημείο κανονισμού |
| 16.8571 | 15 | 7 | 19 | στοιχεία βολής |
| 17.2222 | 16 | 9 | 19 | ύψους διάρρηξης |
| 5.3333 | 8 | 3 | 8 | χαρακτηριστικά σημεία |

Table 3 Collocations that have been also found as nested.

The method is quite simple in the way it works. It is semi-automatic in the sense that it needs the domain-expert and the lexicographer. We believe that this is necessary in order to provide accuracy and completeness to a high degree. However the domain expert and the lexicographer do not have to look (unless really needed) on the corpus itself to obtain the information for the LSP dictionary updating, which of course is a considerable gain on time. We have yet not applied an evaluation measure to judge the results as for the correctness and completeness of information gained. This is a subject still to be done. Another matter is the stemmer. It is not easy to decide whether to use one or not. If yes, words having the same thema would count as one (as they should be in most cases). However there are cases where this should not happen, like with the words "*παρατηρητής*" ("observer") and "*παρατηρητές*" ("observers"). These two words in many cases need to stay as they are found, since they are often used to indicate different meanings in different collocations.

**4 Summary**

In this paper we presented the incorporation of the *C-value* method for the extraction of collocations to dictionary updating. *C-value* offers to this since it focuses on nested collocations, the type of collocations we look at in order to obtain our dictionary new information. The method makes the process faster since we actually look at the extracted collocation list instead of the whole corpus. It is semi-automatic since the

final decision on which information should update the dictionary is taken by the domain expert and the lexicographer.

Regarding future work we first are to apply the method to other languages starting with English, but Turkish, Arabic and Hebrew as well. Should things be working as expected, we will move to the application of the method to multilingual corpora (including parallel) for the updating of multilingual dictionaries.

**References**


Ananiadou S, McNaught J 1995 Terms are not alone: term choice and choice terms. *Journal of Aslib Proceedings* 47(2): 47-60.

Choueka Y, Klein T Neuwitz E 1983 Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of Literary and Linguistic Computing* 4:34-38.

Church K W, Hanks P 1990 Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16:16-29.

Fano R M 1961. *In Transmission of Information: a statistical theory of communications*, New York, M.I.T. Press.

Frantzi K T, Ananiadou S, Mima H 2000 Automatic Recognition of Multi-Word Terms. *International Journal on Digital Libraries* 3(2): 115-130.

Frawley W 1988 Relational models and metascience. In Evens M (ed*) Relational models of the lexicon.* Cambridge, Cambridge University Press, pp 335-372.

Ikehara S, Shirai S, Kawaoka T 1995 Automatic Extraction of Collocations from Very Large Japanese Corpora using N-grams Statistics. *Transactions of Information Processing Society of Japan* 11: 2584-2596

Jones S, Sinclair J 1974 English Lexical Collocations: A Study in Computational Linguistics. *Cahiers de Lexicologie*, 24(1): 15-61.

Kilgarriff A, Tugwell D 2001 WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of Collocation Workshop, ACL 2001, pp 32-38.*

Kim P K, Cho Y K 1993 Indexing Compound Words from Korean Texts using Mutual Information. In *Proceedings of Natural Language Pacific Rim Symposium*, pp 85-92.

Kim Y, Zhang B T, Kim Y T 2001 Collocation Dictionary Optimization Using WordNet and k-Nearest Neighdor Learning. *Journal of Machine Translation* 16: 89-108.

Kita K, Kato Y, Omoto T, Yano Y 1994 A comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria. *Journal of Natural Language Processing* 1: 21-33.

Kjellmer G 1994 *A dictionary of English Collocations*, Oxford: Clarendon Press.

Mima H, Ananiadou S, Neradic G 2001 The ATRACT Workbench: Automatic Term Recognition and Clustering for Terms. In *Lecture Notes in Computer Science, LNAI 2166*, Springer-Verlag, pp 126-133.

Nagao M, and Mori S 1994 A new Method of N-grams Statistics for Lange Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp 611-615.

Palmer F R (ed) 1968 *Selected papers of J.R. Firth*. Harlow: Longman.

Sinclair J, Carter R (eds) 1991 *Corpus, Concordance, Collocation*. Oxford, England, Oxford University Press.

Smadja F 1993 Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19:143-177.