EXPLOITING MINIMAL RESOURCES FOR SUBCATEGORIZATION FRAME ACQUISITION

Katia Kermanidis	UNIVERSITY OF PATRAS
Manolis Maragoudakis	UNIVERSITY OF PATRAS
Nikos Fakotakis	UNIVERSITY OF PATRAS

Περίληψη

Σε αυτή την εργασία χρησιμοποιούμε μια σειρά από στατιστικά φίλτρα (λόγος πιθανοφάνειας, t-σκορ, έλεγχος υποθέσεων, σχετική συχνότητα) σε νέα ελληνικά και αγγλικά σώματα κειμένων για την αυτόματη ανάκτηση άγνωστων πλαισίων υποκατηγοριοποίησης ρημάτων. Επειδή για τις περισσότερες γλώσσες (συμπεριλαμβανομένων και των Νέων Ελληνικών) δεν είναι διαθέσιμοι πάντα πλούσιοι γλωσσολογικοί πόροι, η προεπεξεργασία των κειμένων περιορίζεται στο στάδιο της ενδοπροτασιακής ανίχνευσης μη επικαλυπτόμενων φράσεων. Η πληροφορία σχετικά με τα συμφραζόμενα του ρήματος κωδικοποιείται σαν μια σειρά από φράσεις που προηγούνται και έπονται του ρήματος (ρηματικό περιβάλλον). Κάθε περιβάλλον, καθώς και κάθε υποσύνολό του, αποτελεί ένα υποψήφιο πλαίσιο υποκατηγοριοποίησης. Τα αποτελέσματα που επιτεύχθηκαν είναι συγκρίσιμα και σε πολλές περιπτώσεις καλύτερα από προηγούμενες προσεγγίσεις, ακόμα και προσεγγίσεις που χρησιμοποιούν πιο πλούσιους πόρους.

Λέξεις-κλειδιά

Νέα Ελληνικά, πλαίσια υποκατηγοριοποίησης, στατιστικά φίλτρα

1 Introduction

The detection of the set of syntactic frames, i.e. syntactic entities a certain verb subcategorizes for, is important especially for tasks like parsing and grammar development. Machine-readable dictionaries listing subcategorization frames usually give only expected frames rather than actual ones and are therefore incomplete, or not available for some languages, including Modern Greek (MG). By acquiring frames automatically from corpora, these problems are overcome altogether.

Previous work on learning frames automatically from corpora focuses mainly on English as in Brent (1993), Briscoe and Carroll (1997) and Manning (1993). De Lima (1997) and Eckle and Heid (1996) work on German while Basili et al. (1997) deal with Italian and Zeman and Sarkar (2000) focus on Czech. In most of the above approaches, the input corpus is fully parsed and, if not, only a limited number of frames are learned.

As is the case for the majority of languages, a treebank or a wide coverage syntactic parser are not yet available for MG. Constructing a treebank is expensive and time-demanding. The automatic acquisition of subcategorization information by exploitation of as limited linguistic resources as possible appears to be very challenging.

Contrary to English, which has a more or less fixed-order syntactic structure, in MG the position of the constituents of a sentence is a very weak indicator of their syntactic role. Morphology, on the other hand, is essential for determining verb-argument structure. Based on the above properties of the language, the *environments* of the verbs in the corpus are formed and

counted. The resulting distributions are used as input for several well-known statistical filtering methods we have been experimenting with: relative frequencies, log-likelihood ratio (LLR), t-score, binomial hypothesis testing.

For the present work, pre-processing of the input corpus reaches merely the stage of elementary intrasentential, non-embedded phrase chunking in combination with part-of-speech (pos) and basic morphological tagging. No type of treebank or fully parsed input has been utilized keeping thus the necessary resources to a minimum. At the same time our goal is to learn as many frames as possible, while the complete set of frames for a particular verb is not known to us beforehand but it is detected automatically through the training process. The methodology of Zeman and Sarkar (Zeman and Sarkar, 2000) has been altered in order to better cope with the freedom in MG sentence structure mentioned earlier. We apply our methodology on English as well in order to show its language independence. By *language independence* we mean that the corpus pre-processing methodology requires tools which are feasible to develop in most languages and then by slightly modifying the verb environment formation algorithm one may take into account the special properties of the language in question. For the first time, in this work, comparable results have been obtained by applying the same statistical filters and equivalent pre-processing techniques to the corpora of two distinct languages. A novelty constitutes also the exploration of the effect of automatic pre-processing: taking the phrase structure of one part of the Wall Street Journal directly from the Penn Treebank and creating the phrase structure of the other part by automatically chunking it, we are able to quantitatively explore the effect of automatic vs. manual pre-processing on our task.

Section 2 presents a set of features of Modern Greek that are related to the task. Section 3 describes in detail the pre-processing stage of the Greek and English corpora. The methodology for detecting the environment of a verb and counting the verb and environment occurrences in the data is shown in section 4. All the statistical filtering methods are analyzed in section 5 and their results are evaluated in section 6. The paper concludes in the last section.

2 Relevant aspects of Modern Greek

Regarding morphology, Modern Greek is a highly inflectional language. Nouns, adjectives, articles, participles, ordinal numerals and some types of pronouns are characterized by their case (nominative, genitive, accusative, vocative), gender (masculine or feminine) and number (singular or plural). Gender and number have almost no impact on the detection of subcategorization information, while case is a key feature for the task. Verbs are characterized by their type (main, impersonal), their voice (active and passive), their number and person. The first two of the features affect verb-argument structure significantly.

Concerning sentence structure, a sentence remains grammatically correct and its verb-argument structure remains the same, regardless of the ordering of the phrases it is formed by. Therefore, subcategorization is determined primarily by the morphology rather than the position of the candidate frame.

The following examples demonstrate how the case of the proper noun *Γιάννης* (John) -nominative in a,c and accusative in b-, and the common noun *καρχαρίας* (shark), -nominative in b and accusative in a,c-, determines the arguments of the verb *πιάνω* (to catch).

a. Ο Γιάννης έπιασε τον καρχαρία. (SVO) John caught the shark.
b. Τον Γιάννη έπιασε ο καρχαρίας. (OVS) The shark caught John. (John was caught by the shark).
c. Τον καρχαρία έπιασε ο Γιάννης. (OVS) John caught the shark.

3 Pre-processing

For English we used the Wall Street Journal (WSJ) Corpus as input. For MG we used DELOS (Kermanidis et al., 2002), an automatically annotated MG corpus of approximately five million words and of economic domain. Automatic annotation (pre-processing) of the raw DELOS corpus consisted of the tasks described below in sections 3.1 to 3.3:

3.1 POS and Morphological Tagging

Morphological tagging on DELOS was performed by a morphological analyzer for Modern Greek based on Koskenniemi's two-level morphology model. The processor itself is described in Sgarbas et al. (1995) and utilizes a lexicon (Sgarbas et al., 2000) of more than 60,000 lemmata. The morphological information provided includes part-of-speech tagging for all words, case tagging for nouns, adjectives and pronouns, voice tagging for verbs, type tagging for verbs (distinguishing between personal and impersonal verb types), type tagging for pronouns (distinguishing among relative, interrogative and the rest of the pronouns) and type tagging for conjunctions (distinguishing between coordinating and subordinating conjunctions). Precision and recall values in part-of-speech tagging reach 84-88% and 95-98% respectively. Concerning morphological features that play a key role in the task at hand, case tagging reaches an accuracy exceeding 94%, and voice tagging for verbs 84%.

3.2 Chunking

DELOS has been phrase-analyzed by the phrase-boundary detector (chunker) described in detail in Stamatatos et al. (2000). The chunker is based on very limited linguistic resources, i.e. a small keyword lexicon containing some 450 keywords (articles, pronouns, auxiliary verbs, adverbs, prepositions etc.) and a suffix lexicon of 300 of the most common word suffixes in Modern Greek. In a first stage the boundaries of non-embedded, intrasentential noun (NP), prepositional (PP), verb (VP) and adverbial phrases (ADP) are detected via multi-pass parsing. Smaller phrases are formed in the first passes, while later passes form more complex structures. In Stamatatos et al. (2000) the chunker is reported to achieve a precision of 94.5% and a recall of 89.5% when tested on texts of the MG national newspaper *TO BHMA* (To Vima). In a second stage the head-word of every noun phrase is identified and the phrase inherits its grammatical properties. The head-word identification is based on a set of empirical rules depending on the

case and the pos of the constituents of the phrase. The following example constitutes a sample of the chunker output. The symbol * is used to distinguish the head-word.

VP[*Ανακοινώνεται****] PP[****από την εταιρία****] VP[****ότι ολοκληρώθηκε****] NP[****η *διαδικασία της αυξήσεως του κεφαλαίου της.****]***

VP[*It is being announced****] PP[****by the company****] VP[****that was completed****] NP[****the *process of the increase of its capital.****]***

(It is being announced by the company that the process of increasing its capital has been completed).

As mentioned previously, phrases are non-overlapping. Some of the most characteristic phrase structures are shown in the following table (on the right-hand side of the arrow). On the arrow's left-hand side are the phrase structures they theoretically correspond to. *pr* is a preposition, *n1* and *n2* are constituents of a noun phrase, *v1*, *v2* are constituents of a verb phrase, *adv* is an adverb *con* is a subordinating and *ccon* a coordinating conjunction. *ng* stands for a noun in the genitive case that modifies a previous noun (*n* in the fifth example below).

PP[pr NP[n1 n2]] → PP[pr n1 n2]	*(prepositional phrase structure)*
VP[VP1[v1]S[con VP2[v2]]] → VP1[v1]VP2[con v2]	*(verb+subordinate clause)*
VP[v NP[n]] → VP[v] NP[n]	*(verb+complement phrase structure)*
NP[NP1[n1] ccon NP2[n2]] → NP[n1 ccon n2]	*(coordinate phrase structure)*
NP[n NP1[ng]] → NP[n ng]	*(genitive noun phrase modifier)*
VP[VP1[v] ADP[adv]] →VP[v] ADP[adv]	*(main verb modifier phrase structure)*
NP[ADP[adv] NP1[n]]→ NP[adv n]	*(noun modifier phrase structure)*

Table 1. Description of the corpus phrase structure upon chunking.

3.3 Working with the WSJ

Parts 0001-2454 (appr. 1,2 million words) of the Wall Street Journal from the Penn Treebank (correctly, manually created) have been utilized in our application in a phrase-structure format (provided by the ILK Team at the Tilburg University) rather than the Penn Treebank format in order for English and Modern Greek input data to be comparable. Available morphological information in the WSJ is different from, but equivalent to that in DELOS. For example case tagging is neither existing nor necessary in the WSJ, verb types are distinguished among base forms, gerunds, past participles and modals, pronoun and conjunction types are distinguished.

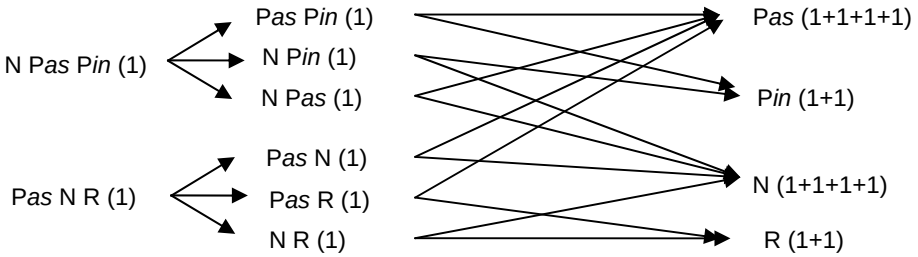
On the other hand, phrases in WSJ parts 2500-6000 (appr. 1,7 million words) have been detected automatically using the Tilburg Memory-based chunker (Daelemans et al., 1999). Using lazy learning, the chunker has been trained on the previous part (0001-2454) of the WSJ and its performance can be found in detail in the above reference.

4 Forming Verb Environments

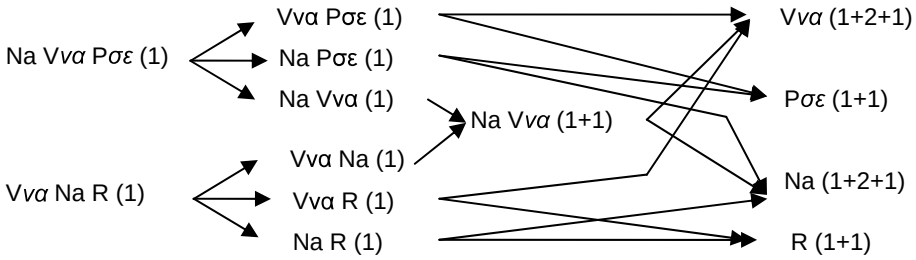
For MG we have carried out a number of experiments concerning the window size of the environment of a verb, i.e. the number of phrases preceding and following the verb. Windows of sizes (-2+3), i.e. two phrases preceding and three phrases following the verb, (-2+2) and (-1+2) were experimented with. For English, which presents a more restricted verb-argument structure regarding the positions of syntactic constituents, we focused on the three phrases following the verb. Unlike MG, in English anything but a subject is highly unlikely to precede the verb.

For both languages almost every environment contains adjuncts. Correct frames are rarely seen isolated in the training data. Therefore, not the entire environment, but one of its subsets is a correct frame of the verb. Large infrequent subsets are likely to contain noise, while smaller, more frequent subsets probably constitute a correct frame. According to the methodology of Zeman and Sarkar (Zeman and Sarkar, 2000), all possible subsets of the above environments were produced and their frequency in the corpus recorded, as shown in Figure 1. Suppose that in the English corpus appear the sentences *We use NP[the thermometer] PP[as an instrument] PP[in this experiment]* and *He uses PP[as a rule] NP[many tools] ADP[every day].* The two encoded environments of *use* are shown in 1a (*N Pas Pin* and *Pas N R). N* stands for noun headword, *R* for adverb, *Pin* for prepositional phrase introduced by *in* and *Pas* for prepositional phrase introduced by *as.* The resulting orderings (subsets) are depicted along with their accumulative counts.

We change the process for counting the subsets of an environment for Modern Greek. Suppose the verb *θέλω* (to want) appears in the MG corpus in the following two sentences: *Θέλει NP[την Τετάρτη] VP[να έρθει] PP[σε μένα]* (He wants [on Wednesday] [to come] [to me]) and *Θέλω VP[να φέρω] NP[το παιδί] ADP[εδώ]* (I want [to bring] [the child] [here]). In Figure 1b the two environments (*Na Vνα Pσε* and *Vνα Na R*) are shown. *Na* stands for a noun headword in the accusative case, *Vνα* for a secondary clause introduced by the conjunction *να , Pσε* for a prepositional phrase introduced by the preposition *σε* and *R* for an adverb. The previous methodology is taken one step further. As verb-argument structure is independent of the ordering of the phrases in a Greek sentence, subsets *Na Vνα* and *Vνα Na* are actually the one and the same as shown in the figure. This is clearly not the case in English, where the ordering of the constituents of the environment has to be maintained and therefore subsets *NPas* and *PasN* are two distinct candidate frames.



(a)



(b)

Fig.1 Forming and counting environments. Numbers next to the environments show their accumulative counts.

A verb appearing in the active as well as the passive voice in the corpus is considered to be two distinct verbs. The same holds for a verb occurring with zero and/ or one and/ or two weak personal pronouns in its verb phrase. Such information affects verb-argument structure significantly.

5 Statistical filtering

The following well-known statistical methods have been applied for the acquisition of the valid frames. Except for the relative frequencies, the rest are based on the same principle: either testing the independence of the distributions of verbs and environments in the data or testing the possibility of a verb co-occurring with an environment, although the latter is not a valid frame.

5.1 Log Likelihood Ratio

Making the hypothesis that the distribution of an environment *e* in the data is independent of the distribution of a verb *v* we can use the log likelihood statistic in order to detect environments highly associated to verbs. According to the parameters introduced by Dunning (1993) the following counts are calculated:

k_1 *the count of a given environment e for a given verb v*

n_1 *the count of a given verb v*

k_2 *the count of environment e with every other verb except for v*

n_2 *the count of every other verb except for verb v*

Using the above values:

$$p_1 = \frac{k_1}{n_1}, \quad p_2 = \frac{k_2}{n_2}, \quad p = \frac{k_1 + k_2}{n_1 + n_2}$$

The log likelihood statistic is then given by:

$$-2\log\lambda = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

where *logL(a,b,c) = clog(a) + (b-c)log(1-a).*

λ is the likelihood ratio for comparing two binomial distributions (Dunning, 1993). The LLR score reflects the difference between the observed (*p1* and *p2*) and the expected expected data (*p*). The greater the score, the more likely it is for the environment to be a valid frame.

5.2 T-score

The t-score statistic, another metric for testing the independence between the two distributions is computed by the following equation, using the definitions from the previous section:

$$T = \frac{p_1 - p_2}{\sqrt{\sigma^2(n_1, p_1) + \sigma^2(n_2, p_2)}}$$

,where

$$\sigma(n, p) = np(1 - p)$$

is the variance of the binomial distribution. Again, if the value of *T* is greater than a threshold value, the candidate frame is labelled as a valid frame of the verb.

5.3 Binomial Hypothesis Testing

Assuming that the data is binomially distributed, each occurrence of a verb *v* is an independent coin flip: as explained in Manning and Schuetze (1999) either an environment *e* may work (heads with probability 1-ε) or may not work (tails with probability ε). In the case of heads, if *e* is a valid frame for *v* it occurs with *v* and correctly indicates being a frame. If it is not a valid frame, it does not occur with *v* and is therefore not misleading. In the case of tails, *v* and *e* co-occur although the latter is not a valid frame. Making the hypothesis that *e* is not a valid frame for *v* (Brent, 1993) we calculate the probability that out of a total of *n* occurrences of *v* in the data, it is seen *m* or more times with *e.* This probability of error is given by:

$$p_E = \sum_{r=m}^{n} \binom{n}{r} \varepsilon^r (1-\varepsilon)^{n-r}$$

If p_E is less than some threshold value, our hypothesis is rejected and *e* is a valid frame for *v.*

5.4 Relative Frequencies

In order to examine the baseline performance for the task at hand we experimented with using a threshold on the relative frequencies (the probability of verbs and frames co-occurring). Using again the previous definitions, a metric that has been used first introduced for the task at hand by Korhonen et al. (2000) is

$$RF = p_1 = k_1 / n_1$$

In a real corpus it is difficult to come across the complete set of the arguments of a verb with a significant frequency. Subsets of this set are much more likely to be found among the environments. These individual subsets, however, co-occur frequently with a large number of

verbs. Provided that the verb appears a significant number of times in the data, the above probability is representative of the most common argument preferences of a verb.

6 Results and Evaluation

To evaluate the approaches we extracted, from parts of each of the three corpora not used for training, all sentences containing one of sixteen verbs [1]. The verbs were chosen randomly, provided that they appeared in the corpus a sufficient number of times (at least 30) and that they presented a variety in syntactic arguments. The environments of these verbs were detected, their subsets were formulated, the occurrences counted and the filtering methods described above were applied to the resulting distributions.

For English we used the COMLEX dictionary (Grishman et al., 1994) as a guide for evaluating the result of the statistical filtering methods. Since a valence dictionary for MG does not exist, it is theoretically impossible to determine with objectivity the entire set of frames that each verb can take. The test sentences were therefore manually tagged by specialists.

We calculated *precision* (the percentage of all the environments labelled by a method as frames which were actually valid frames) and *recall* (the percentage of the actually valid frames which were detected by a method). Accuracy in some domains (such as text classification) is not actually a good metric due to the fact that a classifier may reach high accuracy by simply always predicting the negative class. This problem particularly appears in the present task, where from a vast amount of raw text a large number of invalid frames are produced for every verb.

The results of all the methods on the MG corpus are shown in Table 2. They correspond to the window size (-2+3). After experimenting with the rest of the window sizes (see section 3), with (-2+3) we obtained the best results.

The results for all methods on all the English corpora are shown in Table 3. The higher scores for English could be attributed to the more straightforward and restricted structure of the language. The difference among scores between the two parts of the WSJ shows the impact of correctly vs. automatically pre-processed data. More specifically, errors in the chunking process have a significant effect on the frame acquisition performance.

The probability of the occurrence of a verb with a frame (RF) outperforms the rest of the methods in precision. The number of false positives (instances incorrectly tagged as arguments) with this filter is significantly lower than with the rest. The poorer recall scores indicate that the number of false negatives (instances incorrectly tagged as adjuncts) is substantial and that this filtering method tends to be over-restrictive.

According to Korhonen et al. (2000), although relative frequencies (meaning RF) do not employ any notion of the significance of the observations, their thresholding outperforms the other statistical tests by rejecting low frequency events.

Zeman and Sarkar (2000), Briscoe and Carroll (1997), de Lima (1997), Basili et al. (1997) work on parsed text and so do many of the previous approaches to automatic subcategorization acquisition. They either use a treebank or parsing tools (probabilistic, CFG or other) to reduce the noise in the data. Brent (1993) uses raw text and some basic SF rules and applies Binomial Hypothesis Testing (BHT) on the data, but learns a very limited number of frames. Some of the

approaches, like de Lima (1997) are strongly language-dependent. Moreover, most of them (except for (Zeman and Sarkar, 2000)) pre-suppose knowing all the types of frames in advance.

Precision results with RF are comparable even to previous approaches that employ richer linguistic resources. The best precision score known so far is the one of 88% by Zeman and Sarkar (2000) on fully parsed text. Our recall may be lower than certain recall scores of previous work (as the one reported by Zeman and Sarkar (2000), work on a Czech treebank), but it is mostly comparable to previous results. These numbers are given only as an indication as real comparisons are impossible to make due to the differences in languages, in resources, in the number of learned frames.

DELOS		
Filter	Pr (%)	Re (%)
LLR	50,6	60,3
Ttest	49,9	59,1
BHT	52,4	62,9
RF	70,3	61,1

Table 2. Precision and Recall for DELOS corpora

WSJ	0001-2454		2500-6000	
Filter	Pr (%)	Re (%)	Pr (%)	Re (%)
LLR	54,4	66,2	50,2	60,6
Ttest	51,3	64,1	49,7	57,2
BHT	58,7	71,1	55,3	64,4
RF	77,5	61,7	71,4	58,9

Table 3. Precision and Recall for the English

Conclusion

In this paper we have shown that the automatic acquisition of an unlimited number of subcategorization frames is feasible, even by employing limited linguistic resources for pre-processing the input corpus. Using merely a phrase chunker and keeping in mind the intrinsic properties of Modern Greek and English, we manage to filter out adjuncts with a precision that reaches 78% on unseen data. The frames detected were not known beforehand. As the required pre-processing is elementary, the method can be easily applied to most languages.

Taking advantage of further linguistic knowledge would almost certainly improve the performance for the task at hand. A flexible window size of verb environments, for instance, the boundaries of which are determined by keywords that indicate the end of verb-argument dependence, would be an interesting idea to explore.

As there exist many more languages for which extended subcategorization dictionaries and syntactic treebanks are not available than languages for which such resources are available, the proposal of a system that is able to detect verb frames satisfactorily without the need of such resources is of great significance.

Acknowledgements

The authors would like to thank Sabine Buchholz and Bertjan Busser at Tilburg University for providing the chunked version of WSJ parts 0001-2454 and the Tilburg Memory-based chunker.

Notes

1. For MG the verbs were: αγοράζω (buy), αυξάνω (increase), γνωρίζω (know), δίνω (give), καταλαβαίνω (understand), παράγω (produce), υποθέτω (suppose), χρησιμοποιώ (use) and their passive voice constructions. For English they were: ask, begin, buy, expect, increase, pay, reduce, use and their passive constructions.

References

Basili, Roberto, Pazienza, Maria Teresa, and Vindigni, Michele. 1997. "Corpus-driven Unsupervised Learning of Verb Subcategorization frames". *Proceedings of the Conference of the Italian Association for Artificial Intelligence*. Rome.

Brent, Michael. 1993. "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax". *Computational Linguistics* 19(3). 243-262.

Briscoe, Ted, and Carroll, John. 1997. "Automatic Extraction of Subcategorization from Corpora". *Proceedings of the 5th ANLP Conference, ACL*, 356-363. Washington D.C.

Carroll, John, and Minnen, Guido. 1998. "Can Subcategorization Probabilities Help a Statistical Parser?". *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*. Montreal, Canada.

Daelemans, Walter, Buchholz, Sabine, and Veenstra, Jorn. 1999. "Memory-Based Shallow Parsing". *Proceedings of CONLL-99*, 53-60. Bergen, Norway.

De Lima, Erika. 1997. "Acquiring German Prepositional Subcategorization frames from Corpora". *Proceedings of the 5th Workshop on Very Large Corpora (WVLC-5)*. Beijing, Hong-Kong.

Dunning, Ted. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 19(1). 61-74.

Eckle, Judith, and Heid, Ulrich. 1996. "Extracting raw material for a German subcategorization lexicon from newspaper text". *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'96*. Budapest, Hungary.

Grishman, Ralph, Macleod, Catherine, and Meyers, Adam. 1994. "Comlex syntax: building a computational lexicon". *In Proceedings of the International Conference on Computational Linguistics, COLING-94*, 268-272. Kyoto, Japan.

Kermanidis, Katia, Fakotakis, Nikos, and Kokkinakis Georgios. 2002. "DELOS: An Automatically Tagged Economic Corpus for Modern Greek". *In Proceedings of LREC 2002*, 93-100. Las Palmas de Gran Canaria, Spain.

Korhonen, Anna., Gorrell, Genevieve, and McCarthy, Diana. 2000. "Statistical filtering and subcategorization frame acquisition". *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong.

Manning, Christopher. 1993. "Automatic Acquisition of a Large Subcategorization Dictionary from Corpora". *Proceedings of the 31st Meeting of the ACL*, 235-242. Columbus, Ohio.

Manning, Christopher, and Schutze, Hinrich. 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

Stamatatos, Efstathios, Fakotakis, Nikos, and Kokkinakis, Georgios. 2000. "A Practical Chunker for Unrestricted Text". *Proceedings of the 2nd International Conference of Natural Language Processing (NLP2000)*, 139-150. Patras, Greece.

Sgarbas, Kyriakos, Fakotakis, Nikos, and Kokkinakis, Georgios. 1995. "A PC-KIMMO_Based Morphological Description of Modern Greek". *Literary and Linguistic Computing* 10(3). 189-201.

Sgarbas, Kyriakos, Fakotakis, Nikos, and Kokkinakis, Georgios. 2000. "A Straightforward Approach to Morphological Analysis and Synthesis". *Proceedings of COMLEX 2000*, 31-34. Kato Achaia, Greece.

Zeman, Daniel, and Sarkar, Anoop. 2000. "Learning Verb Subcategorization from Corpora: Counting Frame Subsets". *Proceedings of LREC 2000*, 227-233. Athens, Greece.