Learning Greek Verb Complements: Addressing the Class Imbalance

Katia Kermanidis, Manolis Maragoudakis, Nikos Fakotakis, George Kokkinakis
Wire Communications Laboratory
University of Patras, Rio, 26500, Greece
{kerman,mmarag,fakotaki,gkokkin}@wcl.ee.upatras.gr

Abstract

Imbalanced training sets, where one class is heavily underrepresented compared to the others, have a bad effect on the classification of rare class instances. We apply One-sided Sampling for the first time to a lexical acquisition task (learning verb complements from Modern Greek corpora) to remove redundant and misleading training examples of verb non-dependents and thereby balance our training set. We experiment with well-known learning algorithms to classify new examples. Performance improves up to 22% in recall and 15% in precision after balancing the dataset[1].

1 Introduction

Among the dependents of a verb, *arguments* are key participants in the event described by the verb, while *adjuncts* comprise secondary information concerning the 'setting' of the event (its context, location etc.).

In previous work in automatic complement-adjunct distinction, Buchholz (1998) uses memory-based learning on the part-of-speech tagged and phrase structured part of the Wall Street Journal with a generalization accuracy of 91.6% and she includes verb subcategorization information in her data. Merlo and Leybold (2001) use decision trees to distinguish prepositional arguments from prepositional modifiers. They incorporate semantic verb class, preposition and noun cluster information and reach an accuracy of 86.5% with a training set of 3692 and a test set of 400 instances. Aldezabal et al. (2002) work on Basque. They apply mutual information and Fisher's Exact Test to verb-case pairs (a case is any type of argument) which were obtained from a partially parsed newspaper corpus of 1.3 million words. Evaluation was performed by human tagging of the dependents of ten test verbs inside (55% f-measure) and outside (95% f-measure) the context of the sentence. Many researchers have attempted to distinguish complements from adjuncts as a prerequisite for identifying verb subcategorization frames: Sarkar and Zeman (2000) use a treebank and iteratively reduce the size of the candidate frame to filter out adjuncts. Briscoe and Carroll (1997) and Korhonen et al. (2000) use a grammar and a sophisticated parsing tool for argument-adjunct distinction.

In this paper we address the issue of complement-adjunct distinction in Modern Greek (MG) texts using well-known machine learning techniques (instance based learning, Naïve Bayes, and decision trees) and minimal resources. We make use of input that is automatically annotated only up to the phrase level, where the verb dependents are not identified. Therefore, a significant disproportion between the number of complements and non-complements (adjuncts and non-dependents) arises among the candidates (complements being significantly fewer). This disproportion causes a significant drop in the minority (or positive) class (i.e. complements) prediction accuracy. Henceforth by *adjuncts* we will mean non-complements. The problem of class imbalance has been dealt with in previous work in different ways: oversampling of the minority class until it consists of as many examples as the majority (or negative) class (Japkowicz 2000), undersampling of the majority class (either random or focused), their combination (Ling and Li 1998), the implementation of cost-sensitive classifiers (Domingos 1999), and the ROC convex hull method (Provost and Fawcett 2001).

In general, undersampling the majority class leads to better classifier performances than oversampling the minority class (Chawla et al. 2002). Therefore, we apply *One-sided Sampling* and *Tomek links* (Tomek 1976) to our training data to obtain a more balanced subset of the initial training set by pruning out noisy and redundant instances of the majority class. This approach has been used in the past in several domains such as image processing (Kubat and Matwin 1997), medicine (Laurikkala 2001), text categorization (Lewis and Gale 1994), and we apply it here for the first time to lexical acquisition.

A novel variation in detecting Tomek links in this work is the metric used for calculating the distance between instance vectors. Features in our

[1] This work was supported by the EU Project INSPIRE (IST-2001-32746).

task take exclusively nominal values. We therefore experiment with the *value difference metric* (Stanfill and Waltz 1986) besides the broadly used *Euclidean* distance. The former is more suitable for this type of features, a claim supported by Stanfill and Waltz and also by our experimental results.

2 Modern Greek

Concerning morphology, MG is highly inflectional. The part-of-speech (pos), the grammatical case, and the verb voice are key morphological features for complement detection.

Concerning sentence structure, MG is a 'semi-free' word-order language. The arguments of a verb do not have fixed positions with respect to the verb and are therefore determined primarily by their morphology rather than their position.

Certain semantic verb attributes are also very significant: the verb's copularity, its mode, and whether it is (im)personal. A verb is *copular* when it assigns a quality to its subject. *Mode* is the property that determines the semantic relation between the verb and its subject (whether the latter affects or is affected by the verb action. Although all of these features are normally context-dependent, there are verbs with apriori known values for them. This apriori information is taken into account in our final dataset, as context-dependent semantic information could not be provided automatically, and we tried to keep manual intervention to a minimum.

In MG, verbs can take zero, one or two complements. A complement may be a noun phrase in the accusative or the genitive case, a prepositional phrase or a secondary clause (Klairis and Babiniotis 1999). Often the complements appear within the verb phrase itself in the form of weak personal pronouns. Copular verbs only can take as an argument a noun or adjective in the nominative (predicative). Each of the above features is important but not definitive on its own for complement detection. When combined, however, and including context information of the candidate complement, many cases of ambiguity are correctly resolved. The biggest sources of ambiguity are the accusative noun phrase, which is very often adverbial denoting usually time, and the prepositional phrase introduced by *σε* (to), also often adverbial, denoting usually place.

3 Data Collection

The corpora used in our experiments were:

1. The ILSP/ELEFTHEROTYPIA (Hatzigeorgiu et al. 2000) and ESPRIT 860 (Partners of ESPRIT-291/860 1986) Corpora (a total of 300,000 words). Both these corpora are balanced and manually annotated with complete morphological information. The former also provides adverb type information (temporal, of manner etc.). Further (phrase structure) information is obtained automatically.

2. The DELOS Corpus (Kermanidis et al. 2002) is a collection of economic domain texts of approximately five million words and of varying genre. It has been automatically annotated from the ground up. Morphological tagging on DELOS was performed by the analyzer of Sgarbas et al. (2000). Accuracy in pos tagging reaches 98%. Case and voice tagging reach 94% and 84% accuracy respectively. Further (phrase structure) information is again obtained automatically. DELOS also contains subject-verb-object information limited to nominal and prepositional objects and detected automatically by a shallow parser that reaches 70% precision and recall.

All the corpora have been phrase-analyzed by the chunker described in detail in Stamatatos et al. (2000). Noun (NP), verb (VP), prepositional (PP), adverbial phrases (ADP) and conjunctions (CON) are detected via multi-pass parsing. Precision and recall reach 94.5% and 89.5% respectively. Phrases are non-overlapping. Concerning phrase structure, complements (except for weak personal pronouns) are not included in the verb phrase, nominal modifiers in the genitive case are included within the noun phrase they modify, coordinated simple noun and adverbial phrases are grouped into one phrase.

The next step is empirical headword identification. NP headwords are determined based on the pos and case of the phrase constituents. For VPs, the headword is the main verb or the conjunction if they are introduced by one. For PPs it is the preposition introducing them.

3.1 Data Formation

To take into account the freedom of the language structure, context information of every verb in the corpus focuses on the two phrases preceding and the three phrases following it. Only one out of 200 complements in the corpus appears outside this window. Each of these phrases is in turn the *focus phrase* (the candidate complement or adjunct) and an instance of twenty nine features (28 features plus the class label) is formed for every focus phrase (fp). So a maximum of five instances per verb occurrence are formed. Forming of these instances from a corpus sentence is shown in Figure 1.

The first five features are the verb lemma (*VERB*), its mode (*F1*), whether it is (im)personal (*F2*), its copularity (*F3*), and its voice (*F4*). Two

features encode the presence of a personal pronoun in the accusative (*F5*) or genitive (*F6*) within the VP. For every fp (fps are in bold), apart from the seven features described above, a context window of three phrases preceding the fp and three phrases following it is taken into account. Each of these six phrases (as well as the fp itself) is encoded into a set of three features (a total of twenty one features). These triples appear next in each instance, from the leftmost (-3) to the rightmost phrase (+3). For each feature triple, the first feature is the type of the phrase. The second is the pos of the headword for NPs and ADPs. The third feature for NPs is the case of the headword. For ADPs it is the type of the adverb, if available. If VPs are introduced by a conjunction, the second feature is its type (coordinating/subordinating) and the third is the conjunction itself. Otherwise the second feature is the verb's pos and the third empty. For PPs, the second feature is empty and the third is the preposition.

VP[*](Είναι) NP$_1$[(καλό *παιδί) **NP$_2$[**(ο *Λάμπρος) **CON[**(και) **VP[*](πιστεύει) **PP[*](στο Θεό.)**

(VP[**Is**] NP$_1$[**good boy**] NP$_2$[**the Labros**] CON[**and**] VP[**believes**] PP[**in God.**]) (Labros is a good boy and believes in God.)

VERB	F1 F2 F3	F4	F5	F6	FP	-3	-2	-1	+1	+2	+3	LABEL
είμαι,	O, P, C,	P,	F,	F,	**NP,N,n**,	-,-,-,	-,-,-,	VP,V,-,	NP,N,n,	VP,V,-,	PP,-,σε,	C
είμαι,	O, P, C,	P,	F,	F,	**NP,N,n**,	-,-,-,	VP,V,-,	NP,N,n,	VP,V,-,	PP,-,σε,	-,-,-,	A
πιστεύω,	E, P, NC,	A,	F,	F,	**NP,N,n**,	-,-,-,	-,-,-,	VP,V,-,	NP,N,n,	VP,V,-,	PP,-,σε,	A
πιστεύω,	E, P, NC,	A,	F,	F,	**NP,N,n**,	-,-,-,	VP,V,-,	NP,N,n,	VP,V,-,	PP,-,σε,	-,-,-,	A
πιστεύω,	E, P, NC,	A,	F,	F,	**PP,-,σε**,	NP,N,n,	NP,N,n,	VP,V,-,	-,-,-,	-,-,-,	-,-,-,	C

Figure 1: A sentence is transformed into the 5 labeled instances shown. Words starting with the asterisk (*) are headwords.

The first instance is for the verb *είμαι* and the candidate complement/adjunct is the fp NP$_1$. In the second instance, for the same verb, the candidate complement/adjunct is the fp NP$_2$. There are only two instances for this verb because 1. there are no phrases preceding it, and 2. the third phrase following it (consisting only of the coordinating conjunction) has not much to contribute and is disregarded altogether forcing us to consider the next phrase in the sentence. As the next phrase is a verb phrase that is not introduced by a subordinating conjunction (and therefore cannot be a dependent of the verb *είμαι*), it is also disregarded and no further phrases are tested. In the same way, for the verb *πιστεύω* we have an instance with fp the NP$_1$, an instance with fp the NP$_2$ and one with PP as the fp. We experimented with various window sizes regarding the context of the fp, i.e. [fp], [-1, fp], [-2, fp], [-2, +1], [-3, +3].

The formatting described in the previous section was applied to the ILSP and ESPRIT corpora and to part (approximately 500,000 words) of the DELOS corpus. For the first two corpora, the class of each fp for every created instance was hand-labeled by two linguists by looking up the verb in its context, based on the detailed descriptions for complements and adjuncts by Klairis and Babiniotis (1999). For DELOS, which already contained automatically detected verb-object information to an extent, existing erroneous complement information was manually corrected, while clausal complements were manually detected. The dataset consisted of 63,000 instances. The imbalance ratio is 1:6.3 (one complement instance for every 6.3 adjunct instances).

4 Addressing the Imbalance

From the ratio given above, the complement class is underrepresented compared to the adjunct class in the data. As the number of examples of the majority class increases, the more likely it becomes for the nearest neighbor of a complement to be an adjunct. Therefore, complements are prone to misclassifications. We address this problem with One-sided Sampling, i.e. pruning out redundant adjunct (negative) examples while keeping all the complement (positive) examples. Instances of the majority class can be categorized into four groups (Figure 2): *Noisy* are instances that appear within a cluster of examples of the opposite class, *borderline* are instances close to the boundary region between two classes, *redundant* are instances that can be already described by other examples of the same class and *safe* are instances crucial for determining the class. Instances belonging to one of the three first groups need to be eliminated as they do not contribute to class prediction.

Noisy and borderline examples can be detected using Tomek links: Two examples, *x* and *y*, of opposite classes have a distance of $\delta(x,y)$. This pair of instances constitutes a Tomek link if no other example exists at a smaller distance to *x* or *y* than $\delta(x,y)$.

Redundant instances may be removed by creating a consistent subset of the initial training set. A subset C of training set T is consistent with T, if, when using the nearest neighbor (1-NN) algorithm, it correctly classifies all the instances in T. To this end we start with a subset C consisting of all complement examples and one adjunct example. We

train a learner with C and try to classify the rest of the instances of the initial training set. All misclassified instances are added to C, which is the final reduced dataset.

The exact process of the proposed algorithm is:

1. Let T be the original training set, where the size of the negative examples outnumbers that of the positive examples.

2. Construct a dataset C, containing all positive instances plus one randomly selected negative instance.

3. Classify T with 1-NN using the training examples of C and move all misclassified items to C. C is consistent with T, only smaller.

4. Remove all negative examples participating in Tomek links. The resulting set T_{opt} is used for classification instead of T.4.1 Distance functions

4.1 Distance functions

The distance functions used to determine the instances participating in Tomek links are described in this section.

The most commonly used distance function is the Euclidean distance. One drawback of the Euclidean distance is that it is not very flexible regarding nominal attributes. The *value difference metric (VDM)* is more appropriate for this type of attributes, as it considers two nominal values to be closer if they have more similar classifications, i.e. more similar correlations with the output class. The *VDM* of two values a_x and a_y of a nominal attribute A in two vectors *x* and *y* is estimated as:

$$vdm_A(a_x, a_y) = \sum_{c \in C} \left| \frac{N_{A,a_x,c}}{N_{A,a_x}} - \frac{N_{A,a_y,c}}{N_{A,a_y}} \right|$$

$N_{A,a}$ is the number of times value *a* of attribute *A* was found in the training set, $N_{A,a,c}$ is the number of times value *a* co-occurred with output class c and *C* is the set of class labels.

4.2 The reduced dataset

We used the above distance metrics to detect examples that are safe to remove, and then applied the methodology of the previous section to our data. Figure 3 depicts the reduction in the number of negative instances for both metrics and every fp context window. The more phrases are considered (the higher the vector dimension), the noisier the instances, and the more redundant examples are removed. For small windows, the positive effect of VDM is clear (more redundant examples are detected and removed). As the window size increases, the Euclidean distance becomes smoother (depending on more features) and leads to the removal of as many examples as VDM.

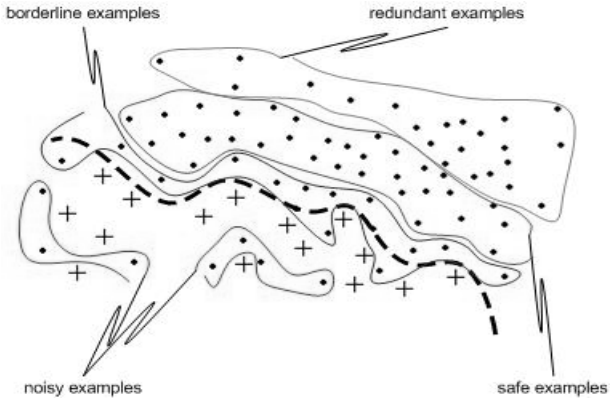

Figure 2: The four groups of negative instances.

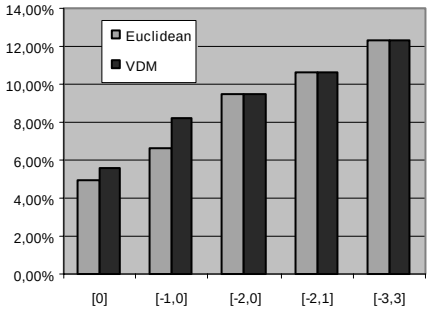

Figure 3: Reduction (%) in the number of negative instances after applying One-sided Sampling.

It is interesting to observe the type of instances which are removed from the initial dataset after balancing. Redundant instances are usually those with as fp headword a punctuation mark, a symbol etc. Such fps could never constitute a complement and appear in the dataset due to errors in the automatic nature of pre-processing. Borderline instances are usually formed by fps that have a syntactically ambiguous headword like a noun in the accusative case, an adjective in the nominative case if the verb is copular, certain prepositional phrases. The following negative instance of the initial dataset (with window [fp]) shows the difference between the two distances.

αντικαθιστώ, E, P, NC, A, F, F, **PP**,-,**σε**, A

This instance appears only as negative throughout the whole dataset. If the verb *αντικαθιστώ* (to replace) were omitted, the remaining instance appears several times in the data as positive with a variety of other verbs. The Euclidean distance between these instances is small, while the VDM is greater, because the verb is a feature with a high correlation to the output class. So the above in-

stance is removed with the Euclidean distance as being borderline, while it remains untouched with VDM.

5 Classifying new instances

For classification we experimented with a set of algorithms that have been broadly used in several domains and their performance is well-known: instance-based learning (IB1), decision trees (an implementation of C4.5 with reduced error pruning) and Naïve Bayes were used to classify new, unseen instances as complements or adjuncts. Unlike previous approaches that test their methodology on only a few new verb examples, we performed 10-fold cross validation on all our data: the dataset (whether initial or reduced) was divided into ten sets of equal size, making sure that the proportion of the examples of the two classes remained the same. For guiding the C4.5 pruning process, one of the ten subsets was used as the held-out validation set.

6 Experimental results

Unlike previous approaches that evaluate their methodology using the accuracy metric, we evaluated classification using precision and recall metrics for every class. a and d are the correctly identified adjuncts and complements respectively, b are the adjuncts which have been misclassified as complements and c are the misclassified complements.

$$pr_A = \frac{a}{a+c},\ re_A = \frac{a}{a+b},\ pr_C = \frac{d}{b+d},\ re_C = \frac{d}{c+d}$$

The f-measure for each class combines the previous two metrics into one:

$$\text{f-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision+recall}}$$

Table 1 shows the results for each classification algorithm and various window sizes using the initial dataset before any attempt is made to reduce its size. The drop in performance of the minority class compared to the majority class is obvious. The scores corresponding to the best f-measure for the complement class are indicated in bold.

By explicitly storing and taking into account every training example, IB1 presents a drop in performance as the window size increases due to sparse data. The performance of C4.5 remains relatively stable, regardless of the size of the instance vector. Naïve Bayes leads to a significant number of adjunct instances being labeled as complements. This is attributed to the fact that the Naïve Bayes learner does not take into account conditional dependencies among features. Given that an instance is a complement, for example, if the fp is an adjective in the nominative case, there is a very high probability in reality that the verb is copular. This dependence is not captured by the Naïve Bayes learner.

		[0]	[-1,0]	[-2,0]	[-2,1]	[-3,3]
Naïve Bayes	Pr_A	91.3	92.5	92.4	92.6	92.9
	Re_A	86.4	83.2	82.1	83.1	82.6
	Pr_C	45.5	43.4	41.8	43.4	**43.1**
	Re_A	57.8	65.6	65.7	66.1	**67.8**
C4.5	Pr_A	91.5	91.4	91.3	91.3	91.5
	Re_A	94.9	95.1	95.2	95.1	95.2
	Pr_C	68.0	68.5	68.7	68.2	**68.9**
	Re_C	54.9	54.4	53.9	53.7	**54.7**
IB1	Pr_A	91.7	92.2	91.6	90.0	87.7
	Re_A	93.7	93.8	92.8	91.6	90.0
	Pr_C	63.8	**65.4**	60.6	52.8	40.5
	Re_C	56.9	**59.8**	56.5	47.9	35.1

Table 1: Results for each algorithm and various fp context window sizes using the initial dataset.

Tables 2 and 3 show the classification results after balancing the dataset using the Euclidean distance and VDM respectively. The increase in f-measure after reducing the dataset is very interesting to observe and depends on the size of the fp context window.

When taking into account the fp only, the highest increase is over 8% in complement class f-measure with the Euclidean distance.

When regarding the context surrounding the fp, the positive impact of balancing the dataset is even stronger. As the fp window size increases, Naïve Bayes performs better, reaching an f-measure of over 60% with [-3,+3] (as opposed to 53.4% prior to balancing). Recall with C4.5 increases by 14% in context [-3,+3] after balancing. Instance-based learning, as mentioned earlier is not helped by a lot of context information and reaches its highest score when considering only one phrase preceding the fp. The increase in complement class precision with IB1 exceeds 12% with VDM. This is the experiment which achieved the highest f-measure (73.7%). Regarding larger context windows and IB1, the removal of the noisy and redundant examples seems to compensate for the noise introduced by the increased number of features in the vector. Increase in recall reaches 22%. As a general remark, instance-based learning performs best when the context surrounding the candidate complement is very restricted (at most one phrase preceding the fp), while Bayesian learning improves its performance as the window increases.

In most of the experiments VDM leads to better results than the Euclidean distance because it is more appropriate for nominal features, especially when the instance vector is small. When larger windows are considered, the two metrics have the

same effect. Minor occasional differences (~0.1%) mirrored in the results are attributed to the 10-fold experimentation.

		[0]	[-1,0]	[-2,0]	[-2,1]	[-3,3]
Naïve Bayes	Pr_A	91.1	92.4	92.8	93.0	93.0
	Re_A	87.4	83.2	82.6	84.6	85.1
	Pr_C	49.0	45.7	46.7	50.3	**51.8**
	Re_A	58.4	67.4	70.5	70.9	**71.3**
C4.5	Pr_A	92.3	92.0	91.7	93.2	92.9
	Re_A	95.1	95.2	95.6	94.6	94.9
	Pr_C	72.4	72.4	74.8	73.6	**73.3**
	Re_C	61.7	60.4	60.1	68.5	**68.8**
IB1	Pr_A	93.0	93.8	93.1	92.1	90.2
	Re_A	94.7	95.5	94.6	93.0	90.5
	Pr_C	71.7	**76.5**	73.0	66.7	55.3
	Re_C	65.4	**69.7**	67.5	68.6	56.7

Table 2: Results for the reduced dataset and the Euclidean distance.

		[0]	[-1,0]	[-2,0]	[-2,1]	[-3,3]
Naïve Bayes	Pr_A	91.0	92.5	92.8	93.0	93.2
	Re_A	87.3	83.1	82.6	84.6	85.4
	Pr_C	49.0	46.5	46.7	50.3	**51.6**
	Re_A	58.6	68.6	70.5	70.9	**71.3**
C4.5	Pr_A	92.0	92.6	91.7	93.2	93.0
	Re_A	95.0	95.2	95.6	94.6	94.8
	Pr_C	71.5	74.3	74.8	73.6	**73.2**
	Re_C	60.1	64.6	60.1	68.5	**68.9**
IB1	Pr_A	92.7	93.8	93.1	93.6	90.2
	Re_A	94.4	95.6	94.6	93.0	90.5
	Pr_C	70.4	**77.5**	73.0	66.7	55.3
	Re_C	64.5	**70.3**	67.5	68.6	56.7

Table 3: Results for the reduced set and VDM.

Apart from the positive impact of One-sided Sampling on predicting positive examples, the tables show its positive (or at least non-negative) impact on predicting negative instances. Non-complement accuracy either increases or remains the same after balancing.

Concerning the resolution of the ambiguities discussed in section 2, three classified examples of the verb *ασκώ* (to exercise) with context environment [-1,fp] follow. The first class label is the true and the second is the predicted class. Example (a) has been classified correctly with and without One-sided Sampling. Examples (b) and (c) are the same instance classified without (b) and with (c) One-sided Sampling. Example (b) is erroneously tagged as an adjunct due to class imbalance. The phrase preceding the fp helps resolve the ambiguity in (a) and (c): usually a punctuation mark before the fp (indicated by the triple *NP,F,-*) separates syntactically the fp from the verb and the fp is unlikely to be a complement.

a. ασκώ, E, P, NC, A, F, F, **PP,-,σε**, NP,F,-, A A
b. ασκώ, E, P, NC, A, F, F, **PP,-,σε**, NP,N,a, C A
c. ασκώ, E, P, NC, A, F, F, **PP,-,σε**, NP,N,a, C C

7 Conclusion

In this paper we describe the positive effect of One-sided Sampling of an imbalanced dataset for the first time on the linguistic task of automatically learning verb complements from Greek text corpora. Unlike traditional One-sided Sampling, we employ the VDM metric and show that it is more appropriate for nominal features. We experiment with various learning algorithms to classify new examples and reach a precision and a recall value of 77.5% and 70.3% respectively, having used only a chunker for preprocessing.

References

I. Aldezabal, M. Aranzabe, A. Atutxa, K. Gojenola and K. Sarasola. 2002. Learning argument/adjunct distinction for Basque. *SIGLEX Workshop of the ACL*, pages 42-50. Philadelphia.

T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. *Proceedings of ANLP 1997*, pages 356-363. Washington D.C.

S. Buchholz. 1998. Distinguishing complements from adjuncts using memory-based learning. *Proceedings of the Workshop on Automated Acquisition of Syntax and Parsing, ESSLLI-98*, pages 41-48. Saarbruecken, Germany.

N. Chawla, K. Bowyer, L. Hall and W.P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321-357. Morgan Kaufmann.

P. Domingos. 1999. Metacost: A general method for making classifiers cost-sensitive. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 155-164. San Diego, CA.

N. Hatzigeorgiu et al. 2000. Design and Implementation of the online ILSP Greek Corpus. *Proceedings of LREC 2000*, pages 1737-1742. Greece.

N. Japkowicz. 2000. The class imbalance problem: significance and strategies. *Proceedings of the International Conference on Artificial Intelligence*. Las Vegas, Nevada.

K. Kermanidis, N. Fakotakis and G. Kokkinakis. 2002. DELOS: An automatically tagged economic corpus for Modern Greek. *Proceedings of LREC 2002*, pages 93-100. Las Palmas de Gran Canaria.

C. Klairis and G. Babiniotis. 1999. *Grammar of Modern Greek. II. The Verb*. (in Greek). Athens: Greek Letters Publications.

A. Korhonen, G. Gorrell and D. McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. *Proceedings of the Joint SIGDAT EMNLP Conference*, pages 199-205. Hong Kong.

M. Kubat and S. Matwin. 1997. Addressing the curse of imbalanced training sets. *Proceedings of ICML 97*, pages 179- 186.

J. Laurikkala. 2001. Improving identification of difficult small classes by balancing class distribution. *Proceedings of the Conference on Artificial Intelligence in Medicine in Europe*, pages 63-66. Portugal.

D. Lewis and W. Gale. 1994. Training text classifiers by uncertainty sampling. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3-12. Dublin.

C. Ling and C. Li. 1998. Data mining for direct marketing problems and solutions. *Proceedings of KDD 98 Conference*. New York, NY.

P. Merlo and M. Leybold. 2001. Automatic distinction of arguments and modifiers: the case of prepositional phrases. *Proceedings of the Workshop on Computational Language Learning*, Toulouse, France.

Partners of ESPRIT-291/860. 1986. *Unification of the word classes of the ESPRIT Project 860*. Internal Report BU-WKL-0376.

F. Provost and T. Fawcett. 2001. Robust classification for imprecise environments. *Machine Learning* 42(3): 203-231.

A. Sarkar and D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. *Proceedings of COLING 2000*, pages 691-697. Saarbruecken, Germany.

K. Sgarbas, N. Fakotakis and G. Kokkinakis. 2000. A straightforward approach to morphological analysis and synthesis. *Proceedings of COMLEX 2000*, pages 31-34. Kato Achaia, Greece.

E. Stamatatos, N. Fakotakis and G. Kokkinakis. 2000. A practical chunker for unrestricted text. *Proceedings of NLP 2000*, pages 139-150. Patras, Greece.

C. Stanfill and D. Waltz. 1986. Toward memory-based reasoning. *Communications of the ACM* 29:1213-1228.

I. Tomek. 1976. Two modifications of CNN. *IEEE Transactions on Systems, Man and Communications, SMC-6*:769-772.