A Unified POS Tagging Architecture and its Application to Greek

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, Stelios Piperidis

Institute for Language and Speech Processing (ILSP),
Artemidos 6 & Epidavrou, 151 25 Maroussi, Greece
{xaris, prokopis, voula, spip}@ilsp.gr

Abstract

This paper proposes a flexible and unified tagging architecture that could be incorporated into a number of applications like information extraction, cross-language information retrieval, term extraction, or summarization, while providing an essential component for subsequent syntactic processing or lexicographical work. A feature-based multi-tiered approach (FBT tagger) is introduced to part-of-speech tagging. FBT is a variant of the well-known transformation based learning paradigm aiming at improving the quality of tagging highly inflective languages such as Greek. Additionally, a large experiment concerning the Greek language is conducted and results are presented for a variety of text genres, including financial reports, newswires, press releases and technical manuals. Finally, the adopted evaluation methodology is discussed.

1. Introduction

Part of Speech (POS) tagging is a well-defined problem, where a suitable morphosyntactic tag is assigned to each word given the context in which it appears. Various methodologies have been proposed making use of linguistic (Karlsson et al., 1995; Oostdijk 1991), statistical (Church 1988; Cutting et al., 1992; Merialdo 1994; Ratnaparkhi 1998) and symbolic learning knowledge (Brill 1992; Brill 1995; Daelemans et al., 1996; Roth et al., 1998). The accuracy reported by most current taggers ranges from 96 to 97% but in the case of highly inflective languages such as Czech, error rate ranges from 20 to 6.2% (Hajic 1999). In this direction, this paper points out some of the difficulties encountered when addressing a tagging exercise concerning the highly inflective Greek language with a rich, structured tagset. One such problem is the large size of the tagset, which amounted to 584 different tags. Another problem is the large number of possible different word forms which leads to a large number of unknown words increasing the work load for guessers or producing a lot of unreliable lexicalized rules. The lack of large, reliable gold tagged corpora for training purposes is another issue that is often underestimated when porting a tagging methodology to other languages.

The proposed platform-independent tagging architecture consists of a tokenizer module, a graphical annotation tool, a feature-based multi-tiered tagger (FBT) - a variant of the well-known transformation based tagger (Brill 1995), a visualization tool providing several views and statistics of the results and a resource administration tool which is responsible for the allocation of the available resources (lexica, rule bases, statistics, lists) enabling the distributed profile of the whole system. The tokenizer module follows the finite-state cascading practice taking into consideration language-specific information concerning, among others, numbers, dates, abbreviations and sentence delimiters. The annotation tool provides all the necessary functionality and automation assisting the annotator(s) in their work while logging difficult cases and idiosyncratic examples that require further decision support. During the human annotation process, a set of handcrafted guidelines for the grammatical annotation of ambiguous cases was used and, when considered appropriate, changed and/or enhanced to cater for idiosyncratic phenomena of the Greek language. The FBT tagger presents a number of extensions to the basic transformation-based rule tagger - TBLT tagger (Brill 1995) which improve its accuracy when trained on small corpora. The main appealing feature is that it allows the treatment of various tagging levels based on the degree of granularity desired, while keeping consistent with the pre-specified tagging encoding scheme. This is accomplished by exploiting the feature structure of the word forms. Finally, the resource administration tool caters for the allocation of all available resources as well as for different tagging parameter settings and user profile logging. Different tagging schemes can be envisaged, all conformant to the PAROLE specifications. Output is provided in a number of encoding schemes, where the default follows the XML standard, enabling Web-based interaction.

2. Compilation and annotation of the corpus

The versions of the tagger we are going to present here were trained on a corpus that was composed of 210 files from different genres of texts, and amounted to a total size of ca. 447K tokens.

Special attention was paid to the overall balance of the corpus which was composed of texts from different domains, ranging from financial newswires to political press conferences, and from interviews to computer hardware tests. They were collected from 17 different online sources.

A finite state tokenizer developed at ILSP was used for sentence splitting and the identification of punctuation and wordforms. Moreover, the tokenizer assigned tags like DATE, DIG(it), ABBR(eviation), etc., to certain special word or multi-word units. Following that tool, a version of the POS tagger by Brill(1993), adapted to Greek and trained on a smaller corpus, was used for the initial annotation of the collection.

Since Greek is a morphologically rich language, the tagset used for the exact description of various morphosyntactic phenomena was very large compared to tagsets used by morphosyntactic annotation schemata for other languages. In the Penn Treebank (Bies et al., 1995), only 36 tags are used. In previous work for Greek, tagsets of 58 tags (Petasis et al., 1999) and 146 tags (Papageorgiou et al. 1995) are reported. Our tagset consists of 584 tags as included in the ILSP-PAROLE tagset (Lambropoulou et al., 1996), which is an adaptation to the Greek language of the PAROLE standard for corpus annotation. Certain tags, allowing for rare cases, such as datives of pronoun forms or numerals, were added to the list during the correction process.

We will present some examples of words and their respective tags in order to explain the large size of the tagset and to present some of the information the tagset is trying to capture. For nouns, information about POS, POS type, gender, number and case is encoded. Thus, for the noun *χρήστης / user*, No(un) is chosen among 13 possible POS values. It is also annotated as C(o)m(mon), Ma(sculine), S(in)g(ular) and N(o)m(inative). All these values are combined in the tag NoCmMaSgNm. Similar features are incorporated in the tags for adjectives and articles. In the case of pronouns like the personal pronoun *εμείς / we* (PnPeMa01PlAcWe) more features are encoded. The first two represent POS (pronoun) and POS type (personal), while the rest deal with gender, person, number, and case. The last feature stands for inflection and the value it is assigned in this example is We(ak). Verbs have the longest tag strings with 10 features that convey information on POS type, finiteness, tense, aspect, voice, number, gender and case, the last two being reserved for passive participles. Certain combinations of tags allow "empty" values. Thus, relative pronouns are annotated with Xx for inflection, since the distinction between strong and weak pronouns applies only to personal pronouns.

For the creation of the gold corpus, two linguists worked in parallel for a period of three months, correcting the output of the tagger. They followed guidelines already set in previous work in corpus annotation at ILSP. There was an attempt to augment, clarify and formalize these instructions. Inter-annotation consistency was addressed as a number of files from the corpus were corrected by both linguists, thus allowing for identification and resolution of discrepancies between the two annotators.

The correction process was facilitated by the use of a graphical tool that was implemented in Tcl/Tk. It consisted of two windows, with the input from the tagger appearing in one of them and the gold corpus in the other. By clicking on the line that contained the error, the annotators were able to build a new tag, or correct a tag with the help of drop down menus that presented them with all possible values for a specific feature. The tool did not allow the insertion of invalid combinations of tags. A morphological lexicon was incorporated in the annotation tool. In case a word was already stored with a different tag in the lexicon, the user had the option of selecting and inserting the alternative in the gold corpus without building the tag from scratch. Moreover, the users had the option of creating and consulting a personal lexicon with words not covered by the morphological lexicon, or with words not encountered with a particular tag in that lexicon. The tool aided in the classification and storage of difficult and ambiguous cases by providing the users with an interface to a database where problematic tokens together with their context and any comments by the annotators, were logged. The guidelines, in HTML format, were accessible from the help menu.

```
<S>
Την/The       AtDfFeSgAc
ιταλική/italian     AjBaFeSgAc
αντιπροσωπεία/delegacy NoCmFeSgAc
στήριξε/supported VbMnIdPa03SgXxPeAvXx
κυρίως/mainly AdXxBa
η/the  AtDfFeSgNm
ισπανική/spanish    AjBaFeSgNm
,      PUNCT
που/which     PnReFe03SgNmXx
αντιμετώπιζε/faced VbMnIdPa03SgXxIpAvXx
παρόμοιο/similar    AjBaNeSgAc
πρόβλημα/problem    NoCmNeSgAc
</S>
------------------------------------------------------------
<S>
       <tok class='tok' from='1.2.1\1'>
              <orth>Την</orth>
              <disamb>
              <ctag>AtDfFeSgAc</ctag>
              </disamb>
       <tok class='tok' from='1.2.1\5'>
              <orth>ιταλική</orth>
              <disamb>
              <ctag>AjBaFeSgAc</ctag>
              </disamb>
       <tok class='tok' from='1.2.1\13'>
              <orth>αντιπροσωπεία</orth>
              <disamb>
              <ctag>NoCmFeSgAc</ctag>
              </disamb>
       <tok class='tok' from='1.2.1\27'>
              <orth>στήριξε</orth>
              <disamb>
              <ctag>VbMnIdPa03SgXxPeAvXx
              </ctag>
              </disamb>
              ……
```

Figure 1: Sample sentence from the gold corpus and its XML representation

The final gold corpus had a format similar to the one of the sample in Figure 1, where translations of the tokens are included for readability purposes.

3. Tagger training and testing

The gold corpus was first separated in a training and a testing corpus. The latter was approximately 20% of the entire corpus (ca. 90K tokens) and was composed of 30 files, which were selected so that all text genres would be represented in the testing phase. The training corpus was then split in two parts, of

approximately 178K tokens each. Attention was paid so that each part contained only full sentences. The first part was used for eliciting lexical information, while contextual rules were acquired from the second part. A lexicon (henceforth, training lexicon), which contained all words (~25K entries) of the first part of the training corpus, was also compiled. Each entry of the lexicon was followed by all the tags with which it was encountered, the first tag being the most frequent. The average number of tags per entry in the lexicon was 1.16. Using these resources, we tested two different versions of a transformation based tagger. We implemented both training and testing components of these versions in Perl.

```
   Lexical rules
   ς      hassuf      1      NoCmFeSgGe
1265.91147652562
   NoCmFeSgGe      ις      fhassuf      2
NoCmFePlNm 93.5940476190476

   Contextual rules
   PnReFe03SgNmXx      PnReNe03PlNmXx
PREV1OR2OR3TAG NoCmNePlAc 121
   AtDfNeSgGe              AtDfMaSgGe
NEXT1OR2OR3TAG NoPrMaSgGe 66
```

Figure 2: Sample lexical and contextual rules from the training of the TBLT

3.1. TBLT experiment

In the first experiment, we followed the transformation based learning tagger as it is described in Brill (1993). The learning module works as follows. At the beginning, a baseline corpus is created by a module that assigns a tag to each word, based on simple heuristics. Thus, all words that contain letters not belonging to the Greek alphabet are recognized as foreign words. The rest of the words are tagged as proper nouns, if they start with a capital letter, and as common nouns in all other cases. The instance of the corpus that is created is then compared to the gold corpus. All the errors are identified and actions that would reduce the error rate are retrieved. These actions follow certain predefined patterns, like *Tag the word as a noun if its suffix of length x is y…* There are 8 patterns on which the generation of lexical rules is based, while affix length values range from 1 to 6. Each action is applied to the corpus and the one that produces a new instance that resembles most the gold corpus is stored as the first rule of the system. This rule is applied to the corpus, and the process is repeated by comparing the new instance of the corpus with the gold corpus.

682 lexical rules were acquired after a training period of 10 days on a Sun Ultra 30 running Solaris. Although patterns for both suffixes and prefixes were allowed, the latter appeared in only 58 rules, showing that it is suffix patterns that contribute the most to unknown words guessing.

Two examples of rules that scored high appear in Figure 2. The number that follows the rules is their actual score and it reflects the neat error reduction they caused, i.e. the number of cases the rule assigned a correct tag minus the times it assigned the wrong tag. The first rule changes the tag of all words whose suffix of length 1 is ς to NoCmFeSgGe. This rule is too general and its output will be corrected in certain cases by more specific rules that scored lower. For instance, the second rule examines suffixes of length 2 and incorporates an additional condition. It is not applied to all words that ends in *ις* but only to those that have already been tagged NoCmFeSgGe.

The next step involves the acquisition of rules based on contextual clues. The set of lexical rules generated is applied to the unknown words of the second part of the corpus. All "known" words, i.e. all words existing in the training lexicon, are assigned the most frequent tag from that lexicon. The instance of the corpus created is compared to the gold corpus. An estimation of the errors is performed and the action that corrects most errors is applied to the corpus and stored as the first contextual rule. At this stage, actions follow 22 patterns that combine information on word and tag context. The context that the rules examined involved three words to the left or the right of the current word. Contextual rules are allowed to change tags for both known and unknown words.

During the experiment, 875 contextual rules were acquired in a training period of 3.5 days. We present two of the most frequent rules produced by the system in Figure 2. In the first one, the number and the gender of a relative pronoun are changed to neuter and plural respectively, in case a neuter plural noun appears in one of the three positions before the pronoun. On the other hand, the second rule looks to the right of an article and changes its gender to masculine in case a masculine proper noun follows.

3.2. FBT experiment

The first experiment involved a very log period of training as far as lexical rules were concerned, due mainly to the large tagset we have chosen to use. In the second experiment we replaced lexical rules with lexicons of suffix-tag combinations, assuming that the rich inflectional system of Greek would allow us to capture morphological information for unknown words from the suffixes of words, without the time consuming training phase of the lexical rules. Extracting the suffixes from the first part of the training corpus, together with all possible tags for each suffix, was trivial. We decided to store information on suffixes of length equal to, or less than, 6 characters.

Another difference between the two experiments was that the contextual rules for this experiment were acquired in four different training stages, each one dedicated to a particular set of morphosyntactic features. During the first stage, training focused on basic POS. Subclassification of POS tags (common and proper nouns, types of pronouns, main and impersonal verbs etc) was also addressed at this point. Gender was the main issue in the second

training stage. Verbal features, including the person of pronouns, were examined during the third training phase. At the final stage, the training module dealt with agreement features.

The construction of the initial state of the corpus involves more heuristics than in the first experiment. At the beginning of the first stage, an initial state corpus is produced. To achieve this, each word is first checked against a lexicon of words that belong to close categories POS (henceforth CCW), such as particles, articles, pronouns and conjunctions. This lexicon contains 2470 entries and apart from the CCWs that appeared in the training corpus, it incorporates capitalized or accented versions of these words in an effort to capture all possible orthographic appearances of these words.

The training lexicon is searched next. In case this lookup does not yield any results and the word contains capital letters, the tagger decapitalizes it and adds accent to each of the last 3 vowels of the word, each time repeating the lookup in the lexicon. If a form of the word matches, the most frequent tag is assigned to the word.

If all the lookups in the training lexicon fail, the tagger tries to guess the tag of the word from its suffix. It extracts a suffix of length 6, or the longest possible suffix in case the word has less than 6 characters. In case the suffix exists in the suffix lexicon, the tagger assigns the most frequent tag found to the word. Otherwise, it subtracts the first character of the suffix and tries again.

Finally, as a last resort, the three default tags that were used in the previous experiment are assigned. The initial corpus created is then compared to the gold corpus and contextual rules are acquired as described above. These rules apply to both known and unknown words. The only exceptions are CCWs. Since we assume that the CCW lexicon is fairly exhaustive, we do not allow contextual rules to change the tag of such a word to a tag with which this word does not appear in the CCW lexicon.

One of the most common rules for the basic POS training stage was the one in Figure 4. This rule changes the tag of a word from article to personal pronoun in case a word tagged as verb follows. The Nv (= No Value) substrings represent features that have not yet been examined at a particular training stage.

For the rest of the training stages a similar procedure is followed. The gold corpus and all the resources are mapped according to the set of features the tagger is examining. The only difference from the first training stage is in the lookup in the lexicons which is not allowed to "destroy" information gained from previous stages. A masculine pronoun, for example, cannot be re-tagged as a neuter article during the fourth training stage, independently of the frequency of these tags in the training corpus. At this stage, the tagger is only allowed to add information, as regards the number and the case of the pronoun in our example.

Contextual rules

AtDfNvNvNv PnPeNvNvNvNvNv
NEXTTAG VbMnNvNvNvNvNvNvNvNv 462

NmCdNeNvNvNv NmCdMaNvNvNv
NEXT1OR2TAG NoCmMaNvNv 37

Figure 4: Sample contextual rules from the first and second training stages of the FBTL

A rule from the second training stage is also included in Figure 4. This rule changes the gender of a numeral from neuter to masculine in case it is followed by a masculine noun.

For the four training stages, training times and

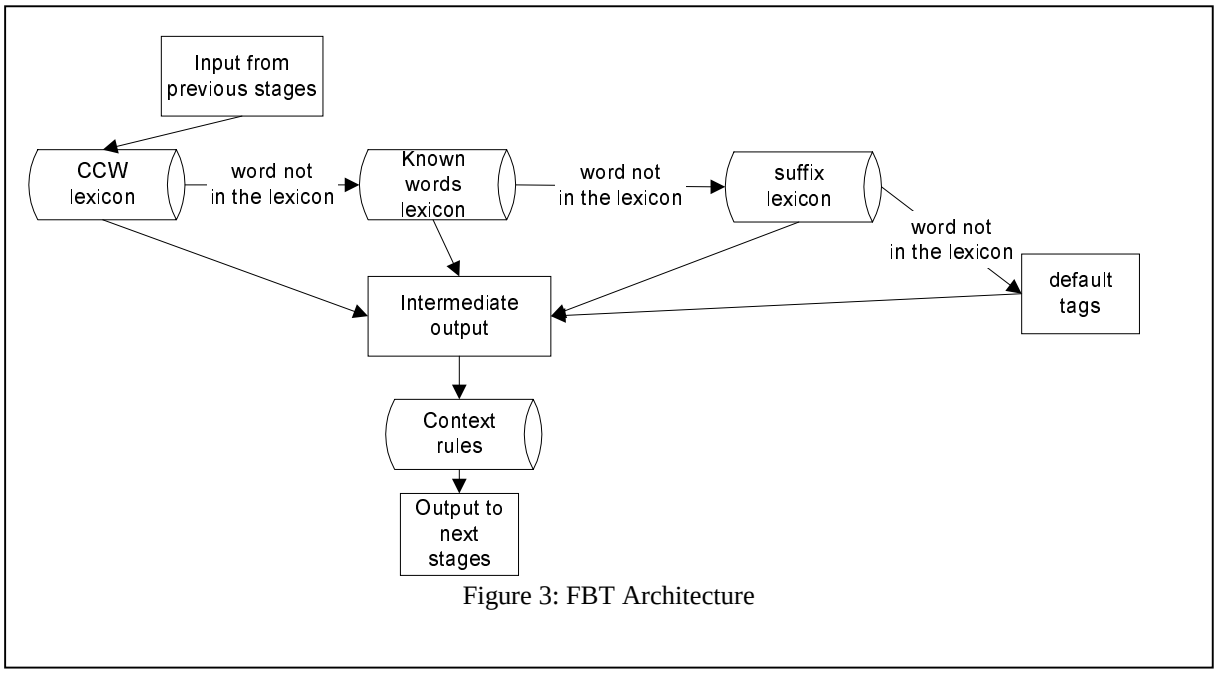

Figure 3: FBT Architecture

number of rules are presented in Table 1.

TRAIN. STAGE	TRAIN. DAYS	RULES
POS-subPOS	0.50	136
Gender	0.48	133
Verbal Features	0.32	53
Agreement	1.46	477
Total	2.76	799

Table 1: Training time and contextual rules for the 4 training stages

In the second experiment, the total training time was reduced to 2.76 days, while better results, presented in the next section, were produced. Furthermore, dividing the training and the applying phase of the tagger in 4 stages allowed us to study the growth of the error in a better way. During the testing phase, the same pipeline architecture, presented in Figure 3 was used.

As far as tagging speed is concerned, it took the tagger 458 seconds to tag the testing corpus file, thus achieving a speed of ca. 200 words / per second on a Pentium III machine.

4. Evaluation

Both versions of the tagger (TBLT and FBT) were tested against a sub-part of the hand-annotated corpus (ca. 90K words) kept aside for evaluation purposes. The performance of each tagger is reported in terms of error rate. It should, however, be made clear that error rate is calculated on the basis of the number of words that have been assigned a POS tag; punctuation, digits, dates, sentence delimiters, etc., recognized by the tokenizer have not been taken into account during accuracy estimation. Global results are listed in Table 2.

TAGGER	BASIC	+GENDER	+VB	ALL
TBLT	4.23	6.26	6.92	10.57
FBT	3.72	5.65	6.30	10.12

Table 2: TBLT and FBT - Global Results

The Transformation-Based Learning Tagger and the Feature Based Tagger produce almost comparable results. Both taggers yield high accuracy rates when only basic category is taken into account. Error rate increases significantly when gender is examined, whereas verbal features do not affect performance seriously. Adding agreement features (case & number) to the previous tags increases error rate considerably. Our methodology of tiered tagging (FBT), however, has proved to be more efficient, rendering better results in each one of the structured disambiguation stages – especially in the basic category.

Although Greek is a highly inflective language, a great level of ambiguity exists between either inflectional forms across different parts of speech or inflectional forms within the same morphological paradigm. The former result in errors concerning basic category, whereas, the latter influence performance regarding gender, verbal, or agreement features. We are going to show how and in what extend ambiguity affected tagging accuracy.

More specifically, the FBT performed well in the basic POS category (error rate 3.72%). Unlike work previously reported (Orphanos et al., 1999), basic POS includes part-of-speech Type. For example, in the case of nouns, their type (Common/Proper) is examined. Impersonal/Personal is encoded in POS type for Verbs, Personal/Possessive for Pronouns, etc. This refined typing clearly affects error rate. Table 3 shows the most common errors concerning POS and POS type.

CORRECT TAG	TAGGER OUTPUT	ERROR DISTRIBUTION
NoCm	AjBa	12.56
AjBa	NoCm	12.45
PnPe	PnPo	7.01
NoPr	NoCm	5.62
AjBa	AdXxBa	3.83
VbIs	VbMn	3.25
NoCm	NoPr	3.10
AdXxBa	AjBa	2.63
PnPo	AtDf	2.52

Table 3: Error distribution relevant to POS and POS type

A closer analysis of the most frequent errors in the basic category shows that the borderline between certain parts-of-speech such as nouns and adjectives, adjectives and adverbs, personal pronouns and possessive pronouns, personal and impersonal verbs, etc., in Greek is not that clear if based on morphology alone. As far as the Noun-Adjective (NoCm-AjBa) pair is concerned, ambiguity arises as they share morphological endings and occur in similar contexts, performing identical syntactic functions (i.e., complement, argument, etc.). Moreover, adjectives are very often used as nouns, too. For example, *πολιτικός* stands for both the adjective *political* and the noun *politician*. Adverbs (AdXxBa) and adjectives (AjBa) have common morphological forms as well. Similarly, certain forms of personal pronouns (PnPe), possessive pronouns (PnPo) and forms of the definite article (AtDf) overlap, and disambiguation of the first two cannot be based on context but, rather, on semantics of their syntactic head (i.e., η καταστροφή *του(PnPe)*/his destruction ~> he is destroyed vs. το αυτοκίνητό *του*(PnPo)/his car ~> he owns the car). Of course, a simplification could be made so as to have these pronouns being always recognized as of type Possessive, but we opted for a more linguistically oriented tagging. Regarding verbs, distinction between personal and impersonal (VbMn and VbIs, respectively) is not based on morphological variety. Finally, ambiguity among CCWs is a difficult task

for by the tagger, since it involves syntactic, apart from morphological, processing. This is, for instance, the case of *γιατί* which is used either as an adverb (AdXxBa) meaning *why*, or as a conjunction (CjSb) with the sense of *because*. Another example is the grammatical word *που/which/who/that/,* being either a personal pronoun or an adverb according to whether it corresponds to a noun phrase or a prepositional phrase respectively.

When gender is added, error rate increases to 5.65%. The most frequent errors regarding gender are shown in Table 4. Apparently, errors occur in ambiguous morphological forms, such as masculine (accusative) and neuter (nominative and accusative) across different parts-of-speech (adjective, definite article, noun). Moreover, masculine and feminine personal and possessive pronouns denoting 1st and 2nd person (εγώ/I, εμένα/me, μου/me/my, με/me, μας/our/us) are not morphologically distinguished even though tagset design imposes the assignment of a value for gender. The same applies to the ambiguous word *που/who/that/which*, which is tagged as PnRe in most contexts. Only if the pronoun's antecedent exists in a three-tokens vicinity, is the tagger able to disambiguate its gender correctly.

CORRECT TAG	TAGGER OUTPUT	ERROR DISTRIBUTION
AtDfMa	AtDfNe	3.90
PnPeFe	PnPeMa	3.73
PnPoFe	PnPoMa	1.76
AjBaMa	AjBaNe	1.76
AtDfNe	AtDfMa	1.68
NoCmFe	NoCmMa	1.49
NoCmMa	NoCmNe	1.40
AjBaNe	AjBaMa	1.40
AtDfNe	AtDfFe	1.25
PnReNe	PnReFe	1.01
…	…	…
		100.00

Table 4: Error distribution relevant to Gender

When verb features (i.e., finiteness, tense, aspect, and voice) are added to the previous estimation, error rate increases to 6.30%. Most errors (25.16%) in verbal features concern number (see Table 5) and occur consistently in the 3rd singular and the 3rd plural of the verb *είμαι/to be*. These forms are the same in Greek, and contextual clues are not always sufficient for disambiguation, especially when ellipsis or coordinated structures are involved. The remaining errors (in Tense+Aspect, Finiteness, or Aspect) are due to verbs with a unique stem for both perfective and imperfective. For example, *θα κάνω* is used for both I *shall do* (imperfective) and *I shall be doing* (perfective), whereas the aspect of most other verbs is shown by means of morphological variation which is not present here. The person of pronouns is addressed together with verbal features (see Appendix , Table 11)

FEATURE	ERROR DISTRIBUTION
Number	25.16
Tense + Aspect	21.29
Finitness	18.49
Aspect	13.98
Voice	4.73
Case (participles)	4.73
Tense + Aspect + Voice	2.37
Other	9.25
Total	100.00

Table 5: Error distribution relevant to Verbal Features

Similarly, when agreement features are taken into account, global error rate increases to 10.12%. It should be mentioned, however, that at this stage of tagging, error rate after lexicon lookup was 15.95%, which was substantially improved when contextual rules were applied. Although word suffixes are, generally, indicative for the part-of-speech of an unknown word, they are not that informative concerning agreement features. There is, indeed, morphological ambiguity between, for example, nominative and accusative in feminine and neuter singular and plural adjectives and nouns, neuter singular and plural nominative and accusative articles, etc. Results concerning the most frequent error distribution regarding agreement features are shown in Table 6.

GOLD	TAGGED	ERROR DISTRIBUTION
NoCmFeSgNm	NoCmFeSgAc	2.71
NoCmNeSgNm	NoCmNeSgAc	2.67
AtDfNeSgNm	AtDfNeSgAc	1.94
NoCmNeSgAc	NoCmNeSgNm	1.92
AtDfNeSgAc	AtDfNeSgNm	1.85
NoCmNePlNm	NoCmNePlAc	1.67
NoCmFeSgAc	NoCmFeSgNm	1.63
AjBaNeSgNm	AjBaNeSgAc	1.57
AjBaFeSgNm	AjBaFeSgAc	1.20
AjBaNePlNm	AjBaNePlAc	1.18
NoCmFePlNm	NoCmFePlAc	1.14
AtDfNePlNm	AtDfNePlAc	1.01
…	…	…
		100.00

Table 6: Error distribution relevant to Agreement Features

Another point to be made is that there is a strong contribution of CCWs to all the figures listed so far.

This contribution, relevant to all words and all CCWs of the testing corpus, is presented in Table 7. The decomposition of the error rate concerning CCWs per stage, is given in the tables of the Appendix.

	BASIC	+GENDER	+VB	ALL
All words	1.03	2.31	2.31	3.73
CCWs	2.35	5.26	5.26	8.50

Table 7: CCWs error contribution

Finally, we separated the texts of the testing corpus in 5 groups according to the domain they fall in (Technical, Financial, General), according to the medium of transmission (Press-Conferences) or by text form (Dialogues). The tagger performed better on texts from the financial domain, whereas low scores were yielded in Dialogues (Table 8). As far as dialogues are concerned, ellipsis phenomena along with turn-taking information not being taken into account, are the main factors for the tagger's not performing well in part-of-speech and in gender.

TEXT TYPE	BASIC	+GENDER	+VB	AGR
General	3.47	5.31	5.76	9.34
Technical	3.76	5.46	5.84	9.77
Finance	2.96	4.28	4.59	7.83
Press-Conf.	2.24	3.44	4.14	8.92
Dialogues	5.26	8.24	9.77	13.76

Table 8: Error Distribution according to text type

5. Conclusions

In this paper we have presented a variation of the well-known transformation-based learning paradigm aiming at improving the quality of tagging highly inflective languages such as Greek. The presented figures and Tables point out the following conclusions:

- FBT tagger yields a higher performance than the TBLT tagger (see Table 1). This is due to the learning procedure followed in the FBT approach. In incremental learning there is an adequate number of cases that support the decision strength of the context-based rules even in small training corpora. Consequently, the acquired rules are more robust with more disambiguating power than in the TBLT approach.
- FBT tagger's performance is in par with other methods that have been applied to highly inflected languages like Czech or when limited only to POS category discrimination. In the latter case, error rate is reduced to 3.72% resulting in 96.28% accuracy. This discrimination level is sufficient for a respectable number of applications.

However, further work needs to be done in several directions. First of all, the combination of two or more taggers seems to be an appropriate exercise enabling the construction of ensembles of classifiers.

Additionally, the incorporation of suffix probabilistic lexica has shown promising results by contributing to the manipulation of the unknown words while eliminating the need of hard-to-get lexical rules. A possible enhancement could be the exploration of more complex context patterns such as the barrier rules considered in constraint grammar formalism (Samuelsson et al., 1996) that give us more disambiguating power in resolving difficult cases that need long distance context.

6. References

Bies, A., M. Ferguson, K. Katz, and R. MacIntyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*.

Brill, E., 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

Brill, E., 1993. Rule based tagger, version 1.14. Available from http://www.cs.jhu.edu/~brill

Brill, E., 1995. Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging. In *Proceedings of the 3rd Workshop on Very Large Corpora*.

Cutting, D., J. Kupiec, J. Pedersen and P. Sibun, 1992. A practical part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

Church, K. W., 1988. A Stochastic Parts program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing*, Austin, Texas.

Daelemans, W., J. Zavrel, P. Berck and S. Gillis, 1996. *MBT: A Memory-Based Part of Speech Tagger-Generator*. In *Proceedings of the 4th Workshop on Very large Corpora*, Copenhagen, Denmark.

Hajic, J. and B. Hledka, 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich Structured Tagset. In *Proceedings of the 36th Annual Meeting of the ACL – Coling*, Montreal, Canada.

Karlsson, F., A. Voutilainen, J. Heikkila and A. Anttila, 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter.

Labropoulou, P., E. Mantzari, and M. Gavrilidou, 1996. *Lexicon - Morphosyntactic Specifications: Language Specific Instantiation (Greek)*, PP-PAROLE, MLAP 63-386 report.

Merialdo, B., 1994. Tagging English Text with a Probabilistic Model. *Computational linguistics*, 20(2):155-171, 1994.

Oostdijk, N., 1991. *Corpus Linguistic and the Automatic Analysis of English*, Rodopi, Amsterdam, Netherlands.

Orphanos, G. and D. Christodoulakis, 1999. Part-of-Speech Disambiguation and Unknown Word Guessing with Decision Trees. In *Proceedings of*

the 9th Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway.

Papageorgiou, H. and S. Piperidis, 1995. A Rule-based tagger Incorporating Statistics. In *Proceedings of the 16th Annual Meeting of the Department of Linguistics*, Thessaloniki, Greece.

Petasis G., G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and I. Androutsopoulos. (1999). Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques. In *Proceedings of the ECCAI Advanced Course on Artificial Intelligence '99*, Chania, Greece.

Ratnaparkhi, A., 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. Dissertation, University of Pennsylvania.

Roth, D. and D. Zelenko, 1998. Part-of-Speech Tagging Using a Network of Linear Seperators. In *Proceedings of the 36th Annual Meeting of the ACL – Coling*, Montreal, Canada.

Samuelsson, C., P. Tapanainen and A. Voutilainen (1996) Inducing Constraint Grammars. In the Proceedings of the 3rd International Colloquium on Grammatical Inference.

7. Appendix

GOLD	TAGGER	ERROR CONTRIBUTION
PnPe	PnPo	25.30
PnPo	AtDf	9.09
AdXxBa	PnRe	7.77
PnPe	AtDf	6.85
CjSb	AdXxBa	4.61
PnPe	AsPpSp	3.69
PnPo	PnPe	3.16
AtDf	PnPo	3.03
PtNg	AdXxBa	2.90
AtDf	PnPe	2.24
AtId	NmCd	2.11
AdXxBa	AsPpSp	1.98
AsPpSp	PnPe	1.71
PnRe	AdXxBa	1.45
AsPpSp	AdXxBa	1.32
AdXxBa	CjSb	1.32
CjSb	PnRe	1.32
…	…	…
		100.00

Table 9: Error contribution of CCWs relevant to POS and POS type

GOLD	TAGGED	ERROR CONTRIBUTION
AtDfMa	AtDfNe	9.55
PnPeFe	PnPeMa	6.19
AtDfNe	AtDfMa	4.13
AtDfNe	AtDfFe	3.07
AtDfMa	AtDfFe	3.01
PnPoFe	PnPoMa	2.83
PnPeFe	PnPeMa	2.65
PnReNe	PnReFe	2.48
PnReFe	PnReNe	1.71
PnReMa	PnReFe	1.42
PnReFe	PnReMa	1.24
PnPoNe	PnPoMa	1.18
PnReMa	PnReNe	1.18
AtDfFe	AtDfMa	1.06

Table 10: Error contribution of CCWs relevant to Gender

GOLD	TAGGER	ERROR CONTRIBUTION
PnPeMa02	PnPeMa01	0.29
PnPeNe01	PnPeNe03	0.18
PnPeMa01	PnPeMa02	0.12
PnPoMa01	PnPoMa03	0.06
PnPeMa03	PnPeMa01	0.06

Table 11: Error contribution of CCWs relevant to the Person feature in pronouns

GOLD	TAGGER	ERROR CONTRIBUTION
AtDfNeSgNm	AtDfNeSgAc	5.25
AtDfNeSgAc	AtDfNeSgNm	5.03
AtDfNePlNm	AtDfNePlAc	2.74
AtDfNePlAc	AtDfNePlNm	1.39
PnReNe03SgAcXx	PnReNe03SgNmXx	1.24
PnReFe03SgAcXx	PnReFe03SgNmXx	1.09
AtIdFeSgNm	AtIdFeSgAc	1.02
AtIdNeSgNm	AtIdNeSgAc	0.98

Table 12: Error contribution of CCWs relevant to Agreement features