Athens University of Economics and Business

Department of Informatics

AUEB's Greek Part-Of-Speech Tagger

Version 1.0

User's Manual

C. Pappas, P. Malakasiotis and I. Androutsopoulos

November 2008

Contents

1. Introduction

This document explains how to use AUEB's Greek Part-of-Speech Tagger (POS tagger). The tagger attempts to automatically determine the part of speech (e.g., noun, adjective, verb, etc.) of each word occurrence in Greek texts. It can also be used to tag each word occurrence with additional information, such as the gender, number, and case of each noun, the voice, number and tense of each verb etc. The tagger uses a *k*-nearest neighbors (*k*-NN) classifier, which is trained on manually tagged texts by using passive or active learning techniques; in the latter case, the system itself suggests training examples, i.e., words to be manually annotated, from a collection of non-annotated training texts. The current version of the tagger was developed during the final-year undergraduate project of Costas Pappas in the Natural Language Processing Group of the Department of Informatics of the Athens University of Economics and Business (AUEB); it is based on earlier versions of the tagger developed by Prodromos Malakasiotis and Ioannis Chronakis.[1]

The tagger is implemented in Java and it is freely available as open-source software with a GNU General Public License (GPL); please read the accompanying file *COPYING.TXT* for more information on GPL. **Please note that the tagger is a research prototype. It is provided with absolutely no guarantee and absolutely no support**.[2]

The tagger invokes a Greek sentence splitter that was developed by Giorgio Lucarelli. The sentence splitter uses LIBSVM.[3] For convenience, a copy of LIBSVM is included in the tagger's software; please read the file *LIBSVM-COPYRIGHT.txt* for information on LIBSVM's copyright.

2. Installation instructions

You need JDK 6 (or later) or JRE 6 (or later) installed on your computer to use the tagger.[4] To check which version (if any) of JDK or JRE is currently installed on your computer, type the following in a command-line shell:

```
java -version
```

---

[1] The theses are available (in Greek) from http://pages.cs.aueb.gr/nlp/theses.html.

[2] However, please send bug reports to Costas Pappas (costaspappus@gmail.com) and Prodromos Malakasiotis (rulller@aueb.gr).

[3] LIBSVM is available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[4] The JDK and the JRE can be downloaded from http://java.sun.com.

If the command reports version number 1.6 or later, the currently installed JDK or JRE is sufficient; otherwise, you need to install a newer JDK or JRE.

To install the tagger, first uncompress the tagger's distribution file *AUEB_Greek_POS_tagger.tar.gz*.[5] In Windows, you may use a program like WinZip or WinRAR.[6] In Unix/Linux, use the commands:

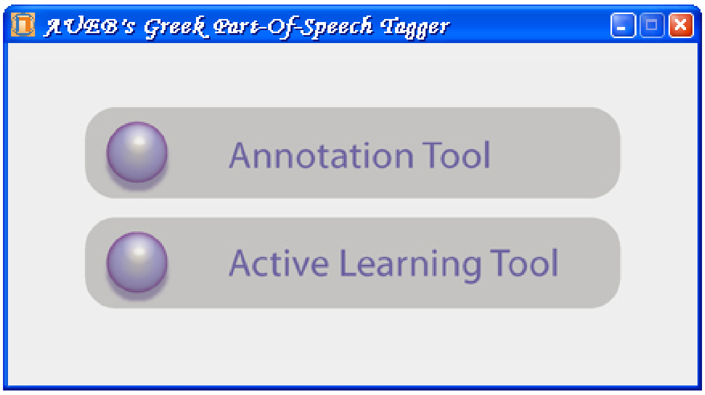
```
> gunzip AUEB_Greek_POS_tagger.tar.gz
> tar xvf AUEB_Greek_POS_tagger.tar
```

The resulting uncompressed directory should contain this manual, two copyright files, the file *POS-Tagger.jar*, as well as the subdirectories *classifier*, *pics*, *texts*, and *src*. The file *POS-Tagger.jar* must always be in the same directory as the subdirectories *classifier*, *pics*, and *texts*. The *src* subdirectory contains the tagger's source code.

To start the tagger double click on *POS-Tagger.jar* or type the following in a command-line prompt:

```
> java –jar POS-Tagger.jar
```

The following window should appear. The options it offers are explained in the following sections.



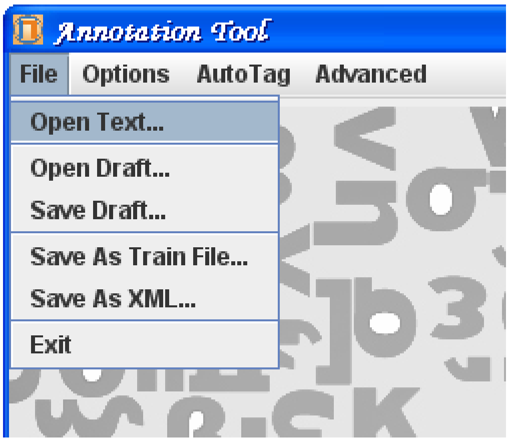
---

[5] The distribution file can be downloaded from http://pages.cs.aueb.gr/nlp/software.html.
[6] See http://www.winzip.com/ and http://www.rarsoft.com/.

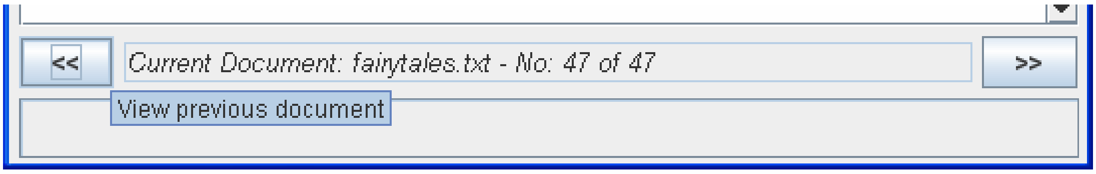
3. Using the pre-trained tagger

The tagger comes pre-trained on Greek newspaper articles. You can use it straight away to automatically tag new Greek newspaper articles or similar texts, as discussed in this section. If you plan to use the tagger with another genre of texts, you may obtain better results by first retraining the tagger on texts of the new genre, as discussed in the following sections.

To use the pre-trained tagger, select the *Annotation Tool* in the initial window (see previous screenshot), and then select *File/Open Text…*, as shown below. A file chooser will appear to allow you to specify the file that contains the text to be tagged. You may also specify a whole directory (folder) of files that contain texts to be tagged. **Please note that in any case the Greek texts must be encoded in UTF-8.**

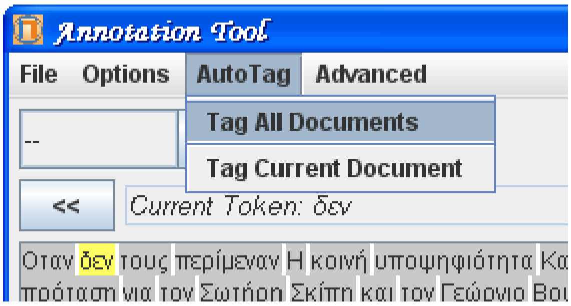


If you specified an entire directory, the text of only one file of that directory will be displayed at any time. You can use the arrows at the bottom of the window to move from one file of the directory to another one (previous or next), as shown below.



Current Document: fairytales.txt - No: 47 of 47

To automatically tag the text of the displayed file, use the *Auto Tag/Tag Current Document* option. If you specified an entire directory, you can use the *Tag All Documents* option to tag all the files of the directory, as shown below.

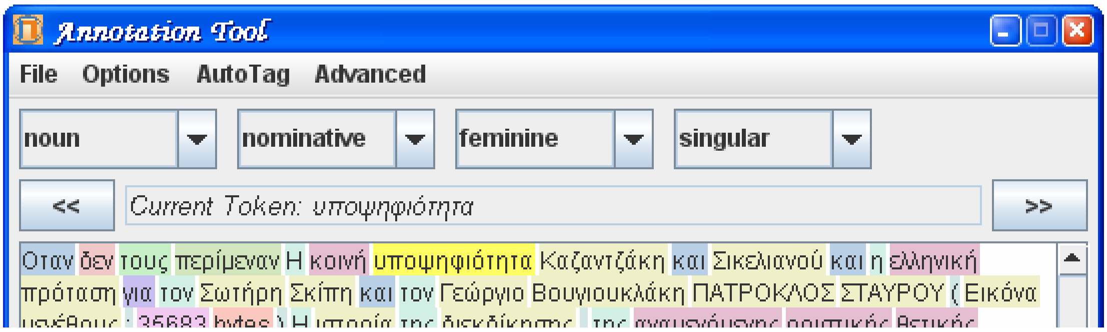


Current Token: δεν

Οταν δεν τους περίμεναν Η κοινή υποψηφιότητα

πρόταση για τον Σωτήρη Σκίπη και τον Γεώργιο

When the automatic tagging is completed, each word will be highlighted with a different color, corresponding to its POS tag. The current word is highlighted in yellow. You may use the arrows at the top of the window to switch to another word (previous or next), as shown below. Alternatively, you can click directly on the word that you want to become the current one.



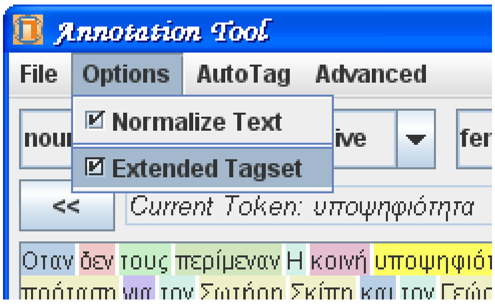
noun | nominative | feminine | singular

Current Token: υποψηφιότητα

Οταν δεν τους περίμεναν Η κοινή υποψηφιότητα Καζαντζάκη και Σικελιανού και η ελληνική

πρόταση για τον Σωτήρη Σκίπη και τον Γεώργιο Βουγιουκλάκη ΠΑΤΡΟΚΛΟΣ ΣΤΑΥΡΟΥ (Εικόνα

The tagger supports two tag sets. If the *Options/Extended Tagset* option is selected (see screenshot below), then, apart from the POS of the corresponding word, each tag will also provide additional information, such as the gender, number, and case of the word, if it is a noun, its voice, number, and tense, if it is a verb etc. Otherwise, each tag will only reflect the POS of the corresponding word.

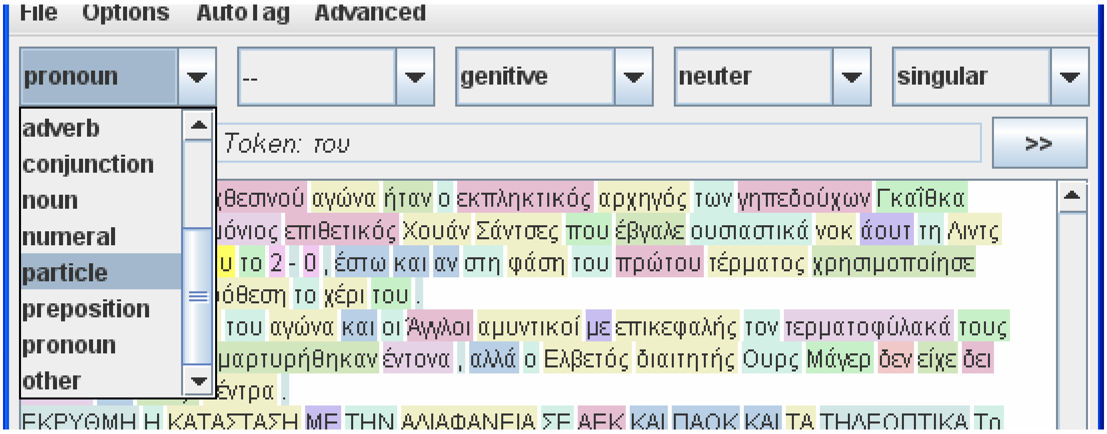


Current Token: υποψηφιότητα

Οταν δεν τους περίμεναν Η κοινή

Although we have been referring to words, the tagger actually assigns a tag to each token; for example, punctuation symbols are treated as separate tokens. The *Normalize Text* option shows the current text in a form where tokens are separated by single spaces.

The information provided by the tag of the current word is shown in the drop-down menus at the top of the window. You may use the drop-down menus to correct the errors of the tagger.



Token: του

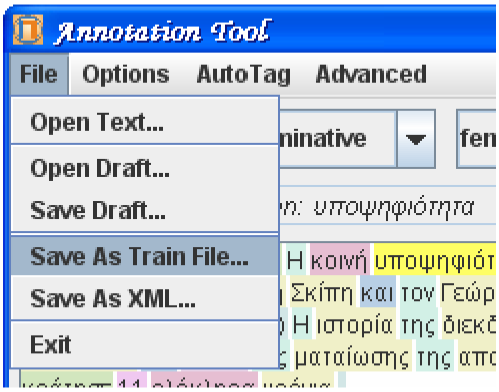
θεσινού αγώνα ήταν ο εκπληκτικός αρχηγός των γηπεδούχων Γκαΐθκα
όνιος επιθετικός Χουάν Σάντσες που έβγαλε ουσιαστικά νοκ άουτ τη Λιντς
υ το 2 - 0 , έστω και αν στη φάση του πρώτου τέρματος χρησιμοποίησε
όθεση το χέρι του .
του αγώνα και οι Άγγλοι αμυντικοί με επικεφαλής τον τερματοφύλακά τους
μαρτυρήθηκαν έντονα , αλλά ο Ελβετός διαιτητής Ουρς Μάγερ δεν είχε δει
έντρα .
ΕΚΡΥΘΜΗ Η ΚΑΤΑΣΤΑΣΗ ΜΕ ΤΗΝ ΑΔΙΑΦΑΝΕΙΑ ΣΕ ΑΕΚ ΚΑΙ ΠΑΟΚ ΚΑΙ ΤΑ ΤΗΛΕΟΠΤΙΚΑ Το

You may also save the (possibly manually corrected) tagged texts in XML by selecting the *File/Save As XML…* option. This is useful when one wants to use other natural language processing tools that require POS tagged texts in XML as input.

4. Retraining the tagger with passive learning

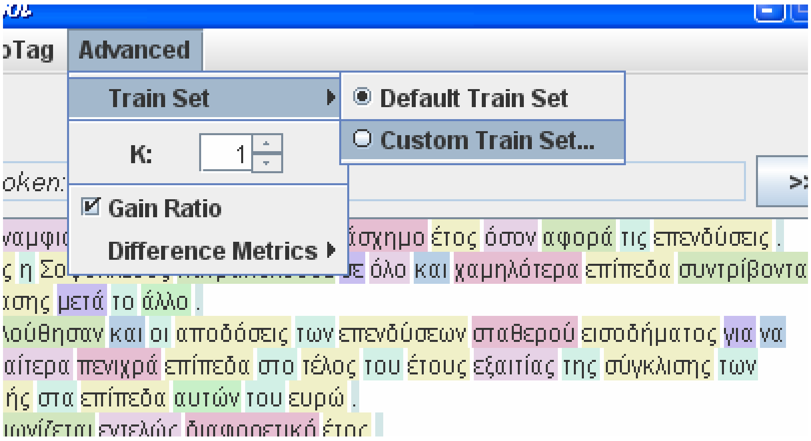
The tagger uses manually tagged texts as training data. The training data are actually stored in a special file format, which stores the tag and context of each manually tagged word occurrence. The default training file contains data from 31,553 manually tagged word occurrences of Greek newspaper articles. As already noted, if you plan to use the tagger with texts of another genre, you may obtain better results if you first retrain the tagger on texts of the new genre. There are two ways to retrain the tagger for texts of a new genre. The simplest way does not use active learning; we call it passive learning, and we discuss it in the remainder of this section. The second way is to use the *Active Learning Tool*, which allows the tagger itself to suggest training examples, i.e., words to be manually annotated; we discuss the use of this tool in the following section.

In passive learning, use the tagger (as already trained) to automatically tag a collection of texts of the new genre (provided as text files in a directory). Then, manually correct the tagger's errors (by using the drop-down menus shown in the previous screenshot), and save the resulting corrected tagged texts as training data. To save a manually corrected set of tagged text files in the special training data format, use the *File/Save As Train File…* option.



If you want to save a temporary copy of a file or a collection of files whose tags you are manually correcting, use the *File/Save Draft…* option. You can then resume the correction at a later time by using the *File/Open Draft…* option.

To instruct the tagger to use a training file (in the special format) that you have created, instead of the default one, select *Advanced/Train Set/Custom Train Set…*, and then select the training file that you want to use. You can also select a folder that contains training files if you wish to train the tagger on all of them.

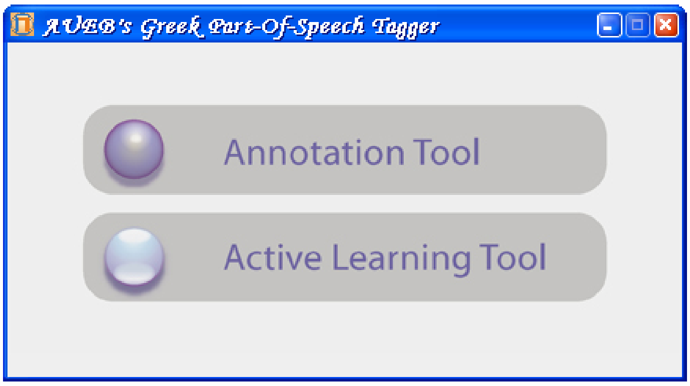

oTag
Advanced
Train Set
● Default Train Set
○ Custom Train Set...
K: 1
☑ Gain Ratio
Difference Metrics
oken:
ίσχημο έτος όσον αφορά τις επενδύσεις .
ε όλο και χαμηλότερα επίπεδα συντρίβοντα
σης μετά το άλλο .
λούθησαν και οι αποδόσεις των επενδύσεων σταθερού εισοδήματος για να
αίτερα πενιχρά επίπεδα στο τέλος του έτους εξαιτίας της σύγκλισης των
ής στα επίπεδα αυτών του ευρώ .

The *Advanced* menu also allows you to adjust other parameters of the tagger, namely the number of nearest neighbors (*k*) that the *k*-NN classifier will consider, whether or not the classifier should use Gain Ratio to weigh the features, and the distance metric that will be used. These parameters are discussed in the theses mentioned in the introduction. It is recommended to keep the default settings.

5. Retraining the tagger with active learning

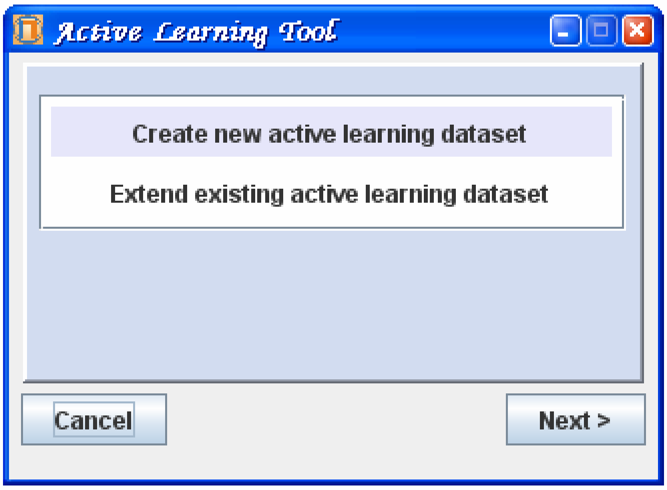
Unlike passive learning, where you need to check and possibly correct the tags assigned by the tagger to the words of the training texts word by word, in active learning you only need to check and correct the tags of particular words selected by the tagger itself from a large pool of untagged texts. Active learning can lead to better classification accuracy with fewer training examples (manually checked tags), i.e., with less manual effort during training. Reducing the number of training examples also increases the classification speed of the *k*-NN classifier.

To train the tagger with active learning, select the *Active Learning Tool* in the tagger's initial window.

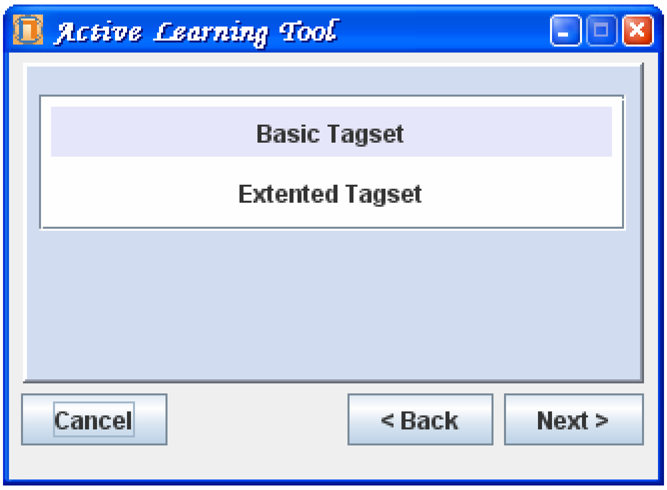


Creating a new active learning dataset

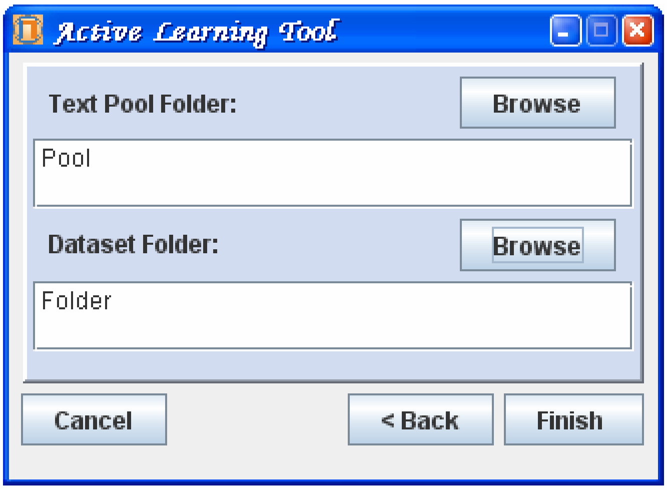
If this is the first time that you use the *Active Learing Tool* for the target genre of texts, select *Create new active learning dataset* in the window that will appear (see below).


Active Learning Tool
Create new active learning dataset
Extend existing active learning dataset
Cancel
Next >

You will then be asked to select between the basic tagset (POS information only) or the extended tagset (POS information, plus gender, number, case of nouns, tense, voice, person of verbs etc.).



Next, you will be asked to specify the *Text Pool Folder* and the *Dataset Folder*. The *Text Pool Folder* must be a directory containing Greek text files **encoded in UTF-8**; this is the collection of texts from which active learning will select examples (words) to be tagged manually. The files of the *Text Pool Folder* will be copied (in a pre-processed form) into the *Dataset Folder*, and the *Text Pool Folder* itself will not be used any furhter. The *Dataset Folder* is a directory (folder) where the new active learning dataset will be stored. The folder must be initially empty.

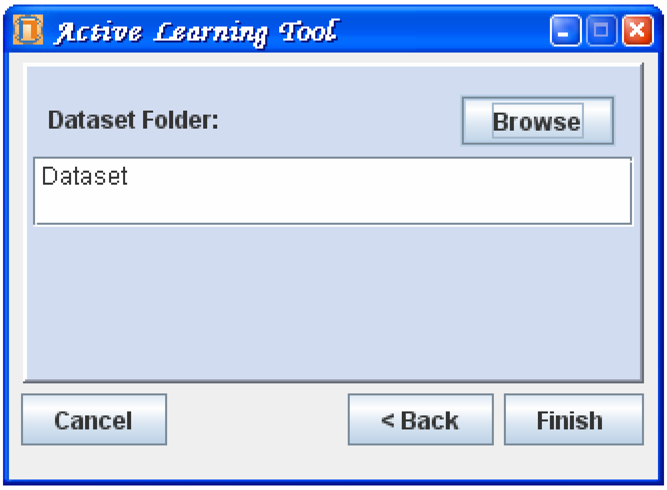


Text Pool Folder: Pool

Dataset Folder: Folder

Extending an existing active learning dataset

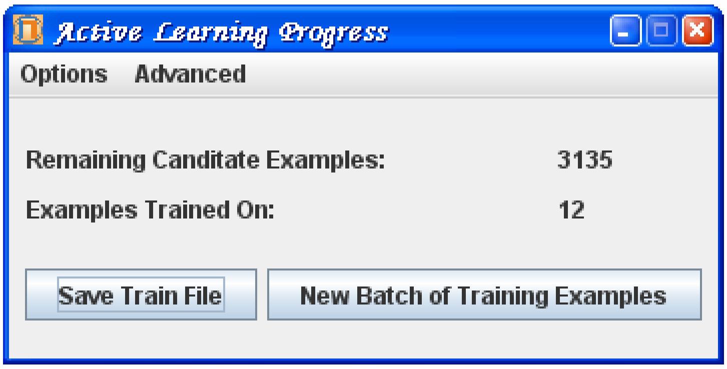
If you have already used the Active Learning Tool in the past to create a training dataset for the target genre of texts and you want to extend the existing training dataset with additional examples, select *Extend existing* *active learning dataset* in the initial window of the *Active Learning Tool*. You will then be asked to specify the directory (folder) where the dataset is stored.



Dataset Folder: Dataset

Adding training examples

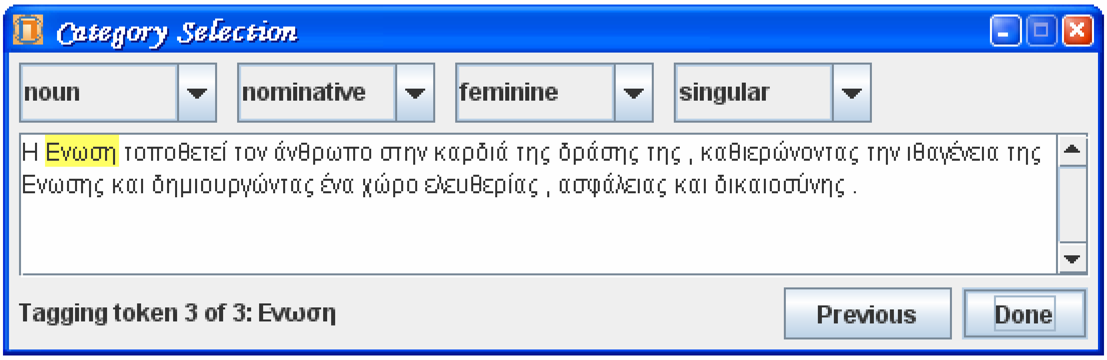
In any case (new or existing dataset), the following window will then appear.



Remaining Canditate Examples: 3135

Examples Trained On: 12

The *Remaining Candidate Examples* field shows the number of words (more precisely, tokens) in the *Text Pool Folder* (more precisely, its copy in the *Dataset Folder*) that have not be used so far as training examples. The *Examples Trained On* field is the number of words (tokens) from *Text Pool Folder* that the system has selected and presented for manual tagging so far. Whenever the *New Batch of Training Examples* button is pressed, the system selects and shows a batch of new training examples (words) from the *Text Pool Folder*, as illustrated below. Use the drop-down menus to assign the correct tag to each training example (each word shown in yellow). Use the arrows to move to the previous or next example of the batch. Press *Done* when you have finished tagging all the words in the batch; this will take you back to the previous window.



noun | nominative | feminine | singular

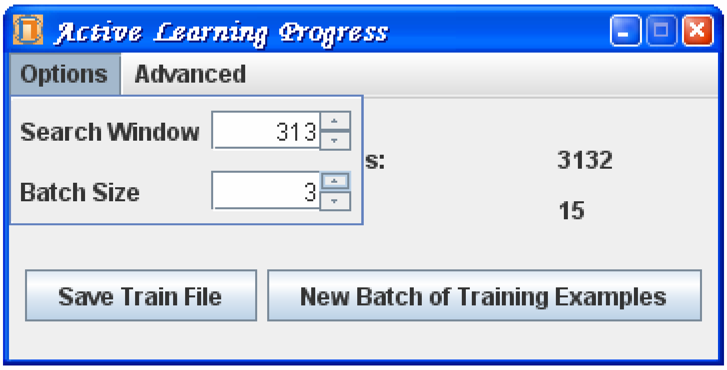
Η Ενωση τοποθετεί τον άνθρωπο στην καρδιά της δράσης της , καθιερώνοντας την ιθαγένεια της Ενωσης και δημιουργώντας ένα χώρο ελευθερίας , ασφάλειας και δικαιοσύνης .

Tagging token 3 of 3: Ενωση

Pressing *Save Train File* in the *Active Learning Progress* window brings up a file chooser, which allows you to specify the location of the resulting training file. The resulting training file has the same format as the files that are produced via the *File/Save As Train File…* option of the *Annotation Tool* when using passive learning (see previous section). You can instruct the *Annotation Tool* to use a training file produced by the *Active Learning Tool* by using the *Advanced/Train Set/Custom Train Set…* option of the *Annotation Tool*.

Other options

The *Options* and the *Advanced* menus of the *Active Learning Progress* window allow you to modify advanced parameters of the *Active Learning Tool*. These parameters are discussed further in the theses mentioned in the introduction. It is recommended to use their default values.



Search Window: 313

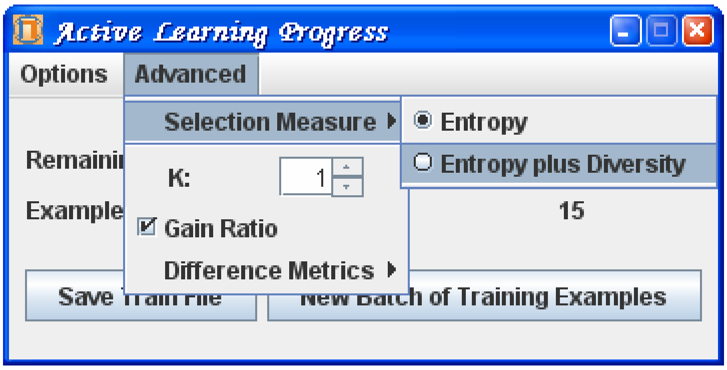
Batch Size: 3

s: 3132

15

The *Search Window* specifies how many candidate examples (words) from the texts in the *Text Pool Folder* are assessed whenever the *New Batch of Training Examples* button is pressed. Ideally, all of the candidate examples would be assessed, but this can be very time consuming. Hence, the system assesses in a circular manner only the next *Search Window* examples (words) of the texts in the *Text Pool Folder*, whenever a new batch is requested.

The *Batch Size* parameter specifies how many examples the system will present for manual annotation whenever it evaluates the next *Search Window* candidate examples of the *Text Pool Folder*. Roughly speaking, the system selects the *Batch Size* examples that it is most uncertain about their tags, given the training it has received so far.



● Entropy

K: 1

☑ Gain Ratio

The *Selection Measure* specifies the measure that is used to assess the usefulness of the candidate training examples.

The *k*, *Gain Ratio*, and *Difference Metrics* options are as in the case of the *Annotation Tool* (see previous section).