NPFL094 - Morphological and Syntactic Analysis
ÚFAL, 24/25 adrirflorez@gmail.com

# Homework 1: Lexicon acquisition

This report is aimed at explaining the technical details of implementation and development, logic and results of the assignment, aimed at acquiring a comprehensive lexicon of Modern Greek language.

## 1. Logic and code

The whole program has been implemented in Python, and can be found in the `code/` directory.

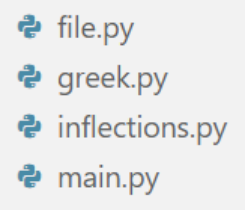


*Figure 1: Code files for lexicon acquisiton, in Python*

Each one of these scripts contain comprehensive documentation for their respective functions. However, a brief description is provided:

- `main.py:` application entry point, to be run directly with only one argument, the filename of the corpus input file, to obtain the corpus.
- `file.py:` utility functions aimed at input from and output to files in different formats (.conllu, .txt, .json).
- `inflections.py:` functions aimed at dictionary creation to associate lemmas to their POS, variants and inflections, as well as their respective annotations.
- `greek.py:` language-specific functions aimed at processing and understanding inflections in Modern Greek, especifically for identifying word stems, inflectional classes and suffixes.

The corpus used as input for the program, from which the lexicon will be acquired, is found in the same directory as all code files. When the program is executed, this is the collection of I/O files in the directory.

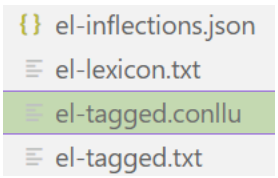


*Figure 2: All files resulting from executing the program*

```
PS C:\Users\adria\Desktop\ÚFAL\NPFL094 - Morphology and Syntax\HW1\code> python main.py el-tagged.conllu
> el-tagged.conllu processed and saved to: el-tagged.txt!
> Inflections saved to: el-inflections.json!
> Lexicon saved to: el-lexicon.txt!
```

*Figure 3: main.py execution output*

## 2. Input and output

The input file `el-tagged.conllu` contains a corpus downloaded for the Universal Dependencies Trees for Modern Greek. It includes sentences and word-per-word tagged information, such as part of speech, annotations and further tags.

el-tagged.conllu

```
# sent_id = gdt-20120309-elwikinews-5160-1
# text = Η Μάντσεστερ Γιουνάιτεντ ηττήθηκε από την Ατλέτικο Μπιλμπάο με σκορ 2:3
1   Η   ο   DET DET Case=Nom|Definite=Def|Gender=Fem|Number=Sing|PronType=Art   2   det _   _
2   Μάντσεστερ  Μάντσεστερ  X   X   Foreign=Yes 4   nsubj   _   _
3   Γιουνάιτεντ Γιουνάιτεντ X   X   Foreign=Yes 2   flat    _   _
4   ηττήθηκε    ηττώμαι VERB    VERB    Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Pas
5   από από ADP ADP _   7   case    _   _
6   την ο   DET DET Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art   7   det _   _
7   Ατλέτικο    Ατλέτικο    X   X   Foreign=Yes 4   obl:agent   _   _
8   Μπιλμπάο    Μπιλμπάο    X   X   Foreign=Yes 7   flat    _   _
9   με  με  ADP ADP _   10  case    _   _
10  σκορ    σκορ    X   X   Foreign=Yes 4   obl _   _
11  2:3 2:3 NUM NUM NumType=Card    10  nmod    _   _
```

*Figure 4: Corpus input file in Modern Greek*

In order to make word-tag processing easier, this file is processed and converted into a simpler `el-tagged.txt` file with maintains the core information needed for obtaining the target lexicon.

el-tagged.txt

```
η   ο   DET DET Case=Nom|Definite=Def|Gender=Fem|Number=Sing|PronType=Art
ηττήθηκε    ηττώμαι VERB    VERB    Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Pass
από από ADP ADP _
την ο   DET DET Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art
με  με  ADP ADP _
χθες    χθες    ADV ADV _
η   ο   DET DET Case=Nom|Definite=Def|Gender=Fem|Number=Sing|PronType=Art
ηττήθηκε    ηττώμαι VERB    VERB    Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Pass
με  με  ADP ADP _
σκορ    σκορ    NOUN    NOUN    Case=Acc|Gender=Neut|Number=Plur
από από ADP ADP _
την ο   DET DET Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art
στα _   _   _   _
σ   σε  ADP ADP _
τα  ο   DET DET Case=Acc|Gender=Neut|Number=Plur
πλαίσια πλαίσιο NOUN    NOUN    Case=Acc|Gender=Neut|Number=Plur
```

*Figure 5: Sample content of the el-tagged.txt file*

The resulting file will be the main input file for our annotations dictionary acquisition, which is aimed at achieving more precise associations between lemmas, their parts of speech, their potential inflections further annotations. This dictionary is to be stored in the `el-inflections.json` file which, although not the target of this homework assignment, will be useful in further iterations of the morphological analyzer.

```
"στο": {
    "pos": "ADP+DET",
    "annotations": [
        "Case=Acc",
        "Gender=Neut",
        "Number=Sing"
    ]
},
"λεπτό": {
    "pos": "NOUN",
    "inflections": {
        "λεπτό": [
            "Case=Acc",
            "Gender=Neut",
            "Number=Sing"
        ],
        "λεπτά": [
            "Case=Acc",
            "Gender=Neut",
            "Number=Plur"
        ]
    }
},
```

```
"ισοφαρίζω": {
    "pos": "VERB",
    "inflections": {
        "ισοφάρισαν": [
            "Aspect=Perf",
            "Mood=Ind",
            "Number=Plur",
            "Person=3",
            "Tense=Past",
            "VerbForm=Fin",
            "Voice=Act"
        ],
        "ισοφαρίσουν": [
            "Aspect=Perf",
            "Mood=Ind",
            "Number=Plur",
            "Person=3",
            "VerbForm=Fin",
            "Voice=Act"
        ]
    }
},
```

*Figure 6: Sample of inflectional information in the dictionary*

Finally, these lemmas, word forms and annotations are parsed and stored into our final `el-lexicon.txt` file, where each line contains the following tab-separated-information: [lemma stem inflectional_class_tag].

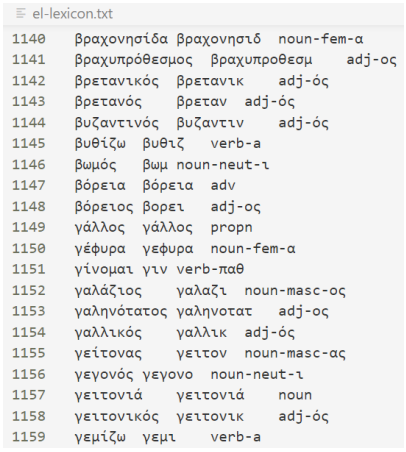


*Figure 7: Sample of the obtained Modern Greek lexicon*

## 3. Heuristics and results

About the aforementioned lexicon file, some observations can be done.

First of all, the results are displayed ordered alphabetically in Greek. The first 400-ish entries of the lexicon correspond to uppercase, proper nouns – as their declensions are interesting enough in terms of morphological analysis.

In order to facilitate the processign of strictly Greek morphology, I have chosen to ignore certain word types, such as:

- punctuation symbols and non-alphabetical characters (incl. numbers),
- foreign words transliterated into Greek alphabet (i.e. *Manchester United* or *Atlético Bilbao*),
- abbreviations and acronyms.

Some workarounds had to be implemented in the case of contractions (that is, joining positional adverbs and determinants into one word), which also present declensions and variations in terms of gender, number and case. They only represent 7 instances in the whole lexicon so, while in the inflections dictionary we can find their detailed information, they were stored as individual lexemes in the lexicon.


*Figure 8: Example of a contraction entry in the lexicon and inflections dictionary, respectively*

Therefore, the word types that shall be found in the lexicon file are:

- Non-inflectional: conjunctions, particles and adverbs of position.
- Inflectional: common and proper nouns, pronouns, verbs, adjectives, numerals and some adverbs.
    - There are some cases of the same word having several potential parts of speech. For simplification, the one containing the most morphological info., the inflectional one, is stored. This is the case of ι.ε. κάτω (positional adverb) and κατώτερα (which refers to lemma κάτω, although as an adjective).
- The aforementioned special case of contractions.

As already mentioned, each lexicon entry follows the structure of `[lemma stem inflectional_class_tag]`. The heuristical analysis followed to achie

- The lemma is the base word form.
- Stem is either the lemma itself or the identified common root with all its inflectional variants. For inflectional words, the Greek accent τόνος is ignored.
- The inflectional class tag can be:
    - For non-inflectional words, their part of speech.
    - For adjectives, verbs, nouns and pronouns, it is their part of speech plus an individual identifier, directly related to their morphological suffix.

The heuristics used to achieve this class tag for inflectional words is strictly based on the words' specific suffixes. This suffix represents how the word will form its declensions in terms of verb conjugation, gender, number or case.

Thankfully, in Greek, there are clearly-defined groups and a list of general, common suffixes that make this task a lot easier:

["αω", "εω", "αμαι", "ωμαι", "ομαι", "ος", "ης", "ας", "μα", "ιο", "α", "η", "ο", "ω"]

These class tags and what morphological suffixes they refer to are explicitly defined in the `greek.py` file, as follows:

```
# ------ VERBS
# https://omilo.com/greek-verbs-for-beginners/
# for verbs, the lemma is first person singular, present
# conjugations A-B1-B2 and passive verbs
VERB_CLASSES = { #
    "b1": ["άω","έω"], # B1 (also with "ώ")
    "b2": ["ώ"], # B2
    "a": ["ω"], # A, first conjugation
    "παθ": ["άμαι", "ώμαι", "ομαι", "ται"], # passive
}
```

*Figure 9: Verb classes*

The main classes for verbs that we consider are first (A) and second (B1, B2) conjugations, as well as passive verbs.

```
# ------ ADJECTIVES
# https://www.greekgrammar.eu/pdffiles/adjectives.pdf
# for adjectives, lemma is masc singular
ADJ_CLASSES = {
    "ός": ["ρός", "νός", "ακός", "ικός", "ός"],    # καλός - good, πικρός - bitter
    "ος": ["ινος", "ιος", "ος"], # αρχαίος - ancient, δερμάτινος - leather
    "ής": ["ής"],   # σταχτής - ash
    "ης": ["ης"], # ζηλιάρης - jealous
    "ύς": ["ύς"],  # βαθύς -- deep
}
```

*Figure 10: Adjective classes*

For adjectives, given that the lemma is always given as masculine singular, we focus on their potential word endings.

```
# ------ NOUNS and PRONOUNS
# https://www.foundalis.com/lan/grknouns.htm
# for gender-variable (pro)nouns, lemma is masc singular
NOUN_CLASSES = {
    # masculine
    "masc-ος": ["ος"],  # άνθρωπος - man
    "masc-ας": ["ας"],  #  πατέρας - father
    "masc-ης": ["ης", "ής"],  # ποιητής - poet
    # neuter
    "neut-μα": ["μα"],  # πρόβλημα - problem
    "neut-ι": ["ι", "ός"], # σπίτι - house, ποτάμι/ός - river
    "neut-ο": ["ο", "ό"],   # βιβλίο - book, βουνό - mount
    # feminine
    "fem-α": ["α"],  # χώρα - country
    "fem-η": ["η", "ή"],  # ψυχή - soul
}
```

*Figure 11: Noun and pronoun classes*

Finally, for nouns (and pronouns, which behave the same when having declensions), we must consider the different suffixes for each one of the genders (masculine, feminine and neuter).

Based on their part of speech, each inflectional word's stem is derived and their potential suffixes identified in order to generalize their potential inflections.

```
3578    στρώμα   στρω     noun-neut-μα
3579    στόλος   στολο    noun-masc-ος
3580    στόχαστρο    στοχαστρ     noun-neut-ο
3581    στόχος   στοχ     noun-masc-ος
3582    συγγένεια    συγγενει     noun-fem-α
```

*Figure 12: Examples of nouns belonging to different noun classes*