A Finite-State Approach to the Computational Morphology of Early Modern Greek

ARISTOMENIS LAMPROPOULOS[1,2], ELENI GALIOTOU[1], IOANNA MANOLESSOU[3], ANGELA RALLI[3]

[1]Department of Informatics
TEI of Athens
Ag. Spyridona, 12210 Egaleo
GREECE

http://www.cs.teiath.gr

[2]Department of Informatics
University of Piraeus
Karaoli & Dimitriou 80,
18534 Piraeus
GREECE

http://students.cs.unipi.gr/~aristomenis

[3]Department of Philology
University of Patras
26500 Rio – Patras
GREECE

Abstract: - We present a finite-state approach to the computational morphology of early Modern Greek that improves the efficiency of searching and accessing to the "Politimo" corpus, which consists of Greek documents printed during the 17th and 18th centuries. Computational morphologies provide users the ability to search documents using only a word root and locate all the corresponding inflected words.

Key-Words: - Computational morphology, Finite state techniques, Natural language processing, Historical document processing

1 Introduction

In recent years, large collections of digitized Greek historical documents were created due to the improvement of Optical Character Recognition Systems. In turn, this fact has given rise to a need for systems that have the ability to manage the content of stored corpus by providing efficient tools for accessing and searching these collections.

The basis of our work is a collection of digitized Greek documents printed during the 17th and 18th centuries. These texts contain a specific morphology, which reflects the evolution of Greek language through the 17th and 18th centuries. Therefore, an efficient accessing system has to take into account the morphology of the language through a systematic linguistic study in order to reveal words that are significant to users, such as historians, linguists etc and describe morpho-phonological rules. The result of that study is the starting point for the construction of computational morphologies which provide users the ability to search documents using only a word root and locate all the corresponding inflected words.

The paper is organized as follows: Sect. 2 describes the framework of "Politimo", while Sect. 3, describes, in relative detail, computational morphologies based on finite state transducers used in our system and the process of analysis and generation of word forms from their morphemes. Sect. 4 presents the tool for the morphological processing of early Modern Greek. Finally, we summarize the paper, draw conclusions and point to related future work in Sect. 5.

2 The framework

The construction of a morphological processor for early Modern Greek is motivated by the need to improve the access to the semantic content of digitized Greek historical and religious documents following the methodology used in the "Politimo" project which is described in the following:

2.1 The "Politimo" project

The "Politimo" project aims at the development and use of innovative techniques of image processing, pattern recognition and natural language processing for the development of an open architecture information system for the processing, management and accessing to the content of valuable collections of historical books and manuscripts.

The aim of the natural language component of the system is the use of advanced natural language processing techniques for the post-processing of the results of the search process and the intelligent and effective searching in the collection such the word-spotting procedure. This approach requires necessarily an extensive linguistic study of the language used during the particular time period.

2.2 The linguistic approach of "Politimo"

Our experimentation is performed on a fragment of the "Politimo" corpus which consists of digitized historical and religious documents printed during the 17th and 18th centuries. The language of the particular time period reflects an early stage of Modern Greek and represents the evolution of the Greek language since it incorporates elements form Ancient, Medieval and Modern Greek. The choice of the particular fragment was made following criteria such as:

- A common set of keywords to facilitate the retrieval process
- Significant size of the corpus
- A language which is close to the spoken language of the particular time period and thus interesting form the point of view of linguistic analysis.

We chose a set of keywords of relevant historical, theological, philological and linguistic interest. These words are mainly nouns of relevant high frequency of appearance which characterize the documents. We also took into account some grammatical elements of particular linguistic interest. These words were further analyzed into their morpheme constituents (roots, inflectional endings, prefixes and suffixes) and morphologically characterized. These morphemes constitute the linguistic basis for the development of our morphological processing tool described in the following sections.

3 Computational Morphology Based on Finite State Transducers

In general, computational morphology [1-7] is concerned with morphological analysis where an inflected word form is analyzed and the base form is returned with additional morpho-syntactic information. This approach calls for a powerful yet simple computational mechanism which is linguistically adequate as well.

Since the emergence of modern linguistics, finite state grammars were put aside as inadequate for the description of linguistic phenomena. Yet, in recent years, an impressive comeback of finite state techniques has been noticed in many NLP application such as tagging, parsing, information extraction etc [8]. Finite state approaches to computational morphology were particularly successful covering a large spectrum of languages including Modern Greek [1, 7]. Therefore, the technology of finite state models, specifically finite state transducers, was a challenging choice for the implementation of our computational morphology which reflects a transitional period in the evolution of the Greek language and has not yet been systematically analyzed.

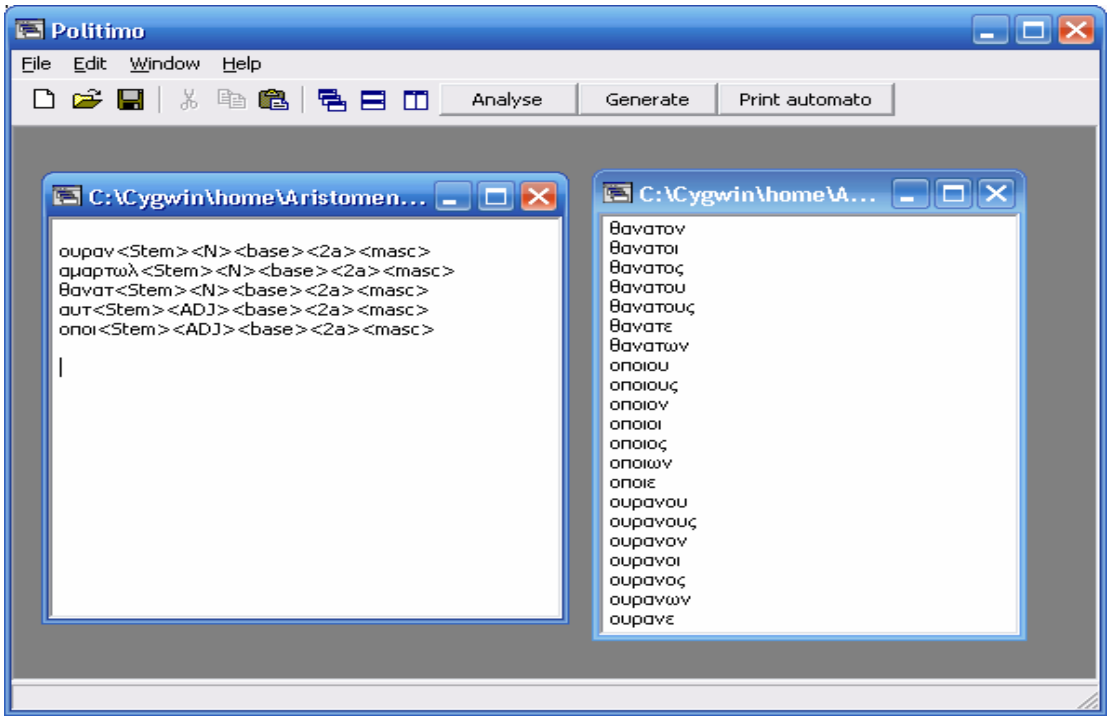

Fig.1: The "Politimo" graphical user interface (lexicon and generated wordforms)

3.1 Finite State Models

A finite state transducer (FST) is basically a finite state automaton where each transition is labeled with a symbol pair rather than a single symbol. Transducers are automata that have transitions labeled with two symbols. One of the symbols represents input, the other - output. Transducers translate (or transduce) strings. In automata theory they are called Mealy's automata. [3], [5].

The finite-state approach to morphology is based on the representation of the relation between the surface forms of a language and their corresponding lexical forms.

This relation can be described as a regular relation using the metalanguage of regular expressions; and, with a suitable compiler, the regular expression source code can be compiled into a finite-state transducer that implements the relation computationally.

As a result, a transducer represents a mapping between a surface form and the lexical form through a sequence of states and arcs from an initial state to a final one.

3.2 The SFST Tool

In our system, we have embedded the SFST tools [3, 4] that provide a programming language for the implementation of finite state transducers which is based on extended regular expressions.

SFST was developed by the Institute for Natural Language Processing, University of Stuttgart. It comprises a compiler, which translates finite state transducer programs into minimized transducers and a wide range of transducer operations similar to commercial platforms such as XFST [5]. Also, it supports UTF-8 character coding which is important for the implementation of Greek computational morphologies.

4 A Tool for the Morphological Processing of Early Modern Greek

We have developed a software tool for creating, manipulating, and applying finite state transducers for the development of computational morphologies which embeds the SFST tools.

We have applied a lexeme-based approach to morphological processing where a word-form is considered as the result of applying rules that alter a word-form or stem in order to produce a new one. An inflectional rule takes a stem, changes it as is required by the rule, and outputs a word-form. Fig.1 and Fig.2 illustrate our tool.

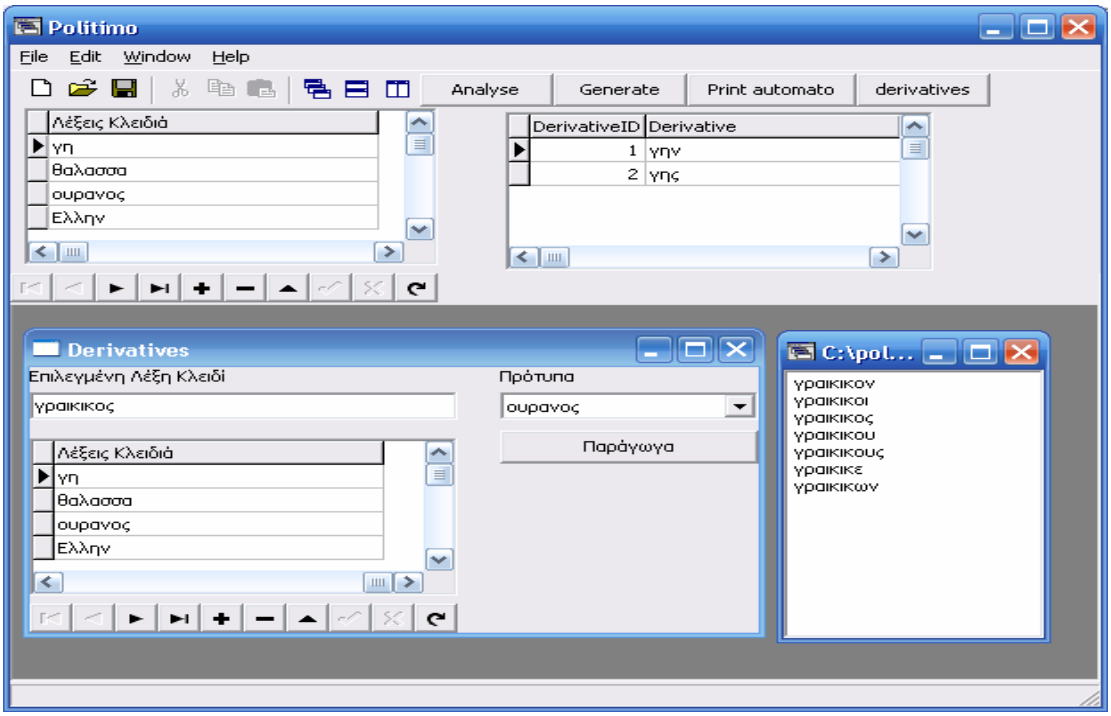

Fig.2: The "Politimo" Tool for the Morphological Processing of Early Modern Greek

Our system consists of two components reflecting two stages of the processing: The first one concerns users with a linguistic background that are familiar with finite state transducers and the SFST

programming language and wish to implement linguistic knowledge. Our system provides a graphical user interface where the experienced user is able to:

- define a lexicon,
- define inflectional classes and assign their representatives
- write morpho-phnological rules
- compile them into an automaton,

as illustrated in Fig.1.

The second component concerns users who use the automaton in order to insert new keywords and enrich the "Politimo" database with new wordforms. Therefore, users of the second component are able to:

- insert a new keyword,
- assign the appropriate inflectional class through the selection of the appropriate representative
- perform a generation of all inflected wordforms according to the function of the selected representative of the inflectional class

as illustrated in Fig. 2.

At this stage of the processing the user needs not be familiar with the finite state formalism.

As a result of this procedure, the system enriches dynamically the list of keywords and inflected wordforms which are used in the word-spotting procedure.

5 Conclusion

In this paper, we have described an attempt – the first to our knowledge - to build a morphological processor for early Modern Greek based on finite state techniques.The aim of the processor is twofold: Improve the access to the content of digitized Greek historical documents and contribute to the theoretical and computational processing of the language of the particular time period. First results of our experimentation have shown that a finite state approach is adequate for the description of morphological phenomena of the particular period. Therefore, future extension of our system will provide an efficient computational tool for the study of the diachronic evolution of the Greek language.

References:

[1] A. Ralli, E. Galiotou, Greek Compounds: "A Challenging Case for the Parsing Techniques of PC-KIMMO v.2", *International Journal of Computational Intelligence*, *vol*. 1, no. 2, pp. 152-162, 2004.

[2] E. Antworth, "PC-KIMMO: A Two-level Processor for Morphological Analysis", *Occasional Publications in Academic Computing no 16*, Summer Institute of Linguistics, Dallas TX, 1990.

[3] H. Schmid, "A Programming Language for Finite State Transducers", *Proc. FSMNLP 2005*, Helsinki, Finland, 2005.

[4] H. Schmid, A. Fitschen, U. Heid, "SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection", *Proc. LREC 2004*, pp. 1263-1266, Lisbon, Portugal, 2004.

[5] K. Beesley, L. Karttunen, "Finite State Morphology", *CSLI Publications*, 2003.

[6] K. Koskenniemi, *Two-level Morphology: A General Computational Model for Wordform Recognition and Production*, Publication No 11, Dept. of General Linguistics, Univeristy of Helsinki

[7] L. Karttunen, KIMMO: "A General Morphological Processor", *Texas Linguistic Forum*, vol. 22, pp.163-186

[8] L. Karttunen, K. Oflazer (eds.) *Special Issue on Finite-State Methods in NLP: Computational Linguistics*, vol. 26, no. 1, 2000.

[9] M. Mohri, "On Some Applications of Finite-State Automata Theory to Natural Language Processing", *Natural Language Engineering*, no. 2, pp. 1-20.

Acknowledgements

This work was co-funded from the E.U. by 75% and from the Hellenic State by 25%, through the Operational Program «Information Society» - Measure 3.3. «Research and Technological Development in the Information Society».