Mediterranean Archaeology and Archaeometry, Vol. 14, No 4, pp. 165-174

THE DEVELOPMENT OF VOCABULARIES OF HISTORICAL PERIOD NAMES FROM WEB ACQUIRED CORPORA7

Maria S. Mouroutsou*[1], Stella Markantonatou[2] and Vasilis Papavasiliou[2]

[1]National and Kapodistrian University of Athens, School of Philosophy, Dept. of Philology
Panepistimiopolis, 15784 Ilissia, Greece
*[2]**ILSP/ "Athena"RIC**, Artemidos 6 & Epidavrou*
GR- 151 25 Maroussi, Greece

Received: 15/12/2013
Accepted: 08/05/2014 *Corresponding author: Maria S. Mouroutsou (msmourou@gmail.com)*

ABSTRACT

Periodization is a universal and very popular system of organizing History (Petras, et al., 2006) by arbitrary dividing time into periods such as "Δικτατορία" (dictatorship) in a way that is specific to places and communities. Structured collections of time period names and timelines are considered very useful in cultural content documentation and temporal information extraction. However, to the best of our knowledge, this is the first report on the systematic collection of period names of Greek History.

New period names are constantly created or left out of use. Aiming to capture this combination of dispersed specificity and constant evolution, we used the Focused Monolingual Crawler (FMC) (Mastropavlos, et al., 2011) and an initial list of 25 "seed-terms" to develop corpora dense in period names with Web retrieved documents. Period names were manually retrieved from the accumulated corpora and were annotated for a set of features, including allomorphs that occurred in the collected corpora and whether the term denoted a fact or a time period or something else as well as for persons, places and other period names related with the term.

The linguistic environments where the terms occurred were identified and some of them were fed to the (FMC) as new "seed-terms". This cycle was repeated for three times and yielded 78 period names with an average of 16 paradigms per term and a corpus consisting of 3020 valid XML documents. Some first observations on the strategies employed by Greek communities to coin time period names are reported.

KEYWORDS: periodization, time period name, Focused Monolingual Crawler, unstructured Web data

1. INTRODUCTION

Periodization is a system of organizing historical information by dividing time into periods. Time period names, a type of named entity in fact, often denote more than simply calendar dates; they implicate a subject, time and place (ISO/CD21127). For example, the Greek period name "χρυσούς αιών" (golden age) encapsulates a place (Athens), time (fifth century), and subject (a flowering of arts and culture, and the height of democracy).

In CIDOC CRM (Doerr 2003; Crofts et al., 2004) the basic notion of a period is defined as follows: "*This class comprises sets of coherent phenomena or cultural manifestations bounded in time and space. It is the social or physical coherence of these phenomena that identify a…period and not the associated spatio-temporal bounds. These bounds are a mere approximation of the actual process of growth, spread and retreat. Consequently, different periods can overlap and coexist in time and space, such as when a nomadic culture exists in the same area as a sedentary culture.*" (Crofts et al., 2004).

In Feinberg et al, (2003) time periods are described more or less as gazetteers that match place names to coordinates. Just like a gazetteer, a time period directory could match time period terms to date ranges, location and other information that characterizes the period.

To these descriptions we would add that time period names are constantly developed and abandoned; the phenomenon is of a dynamic nature.

In the light of the above descriptions of the notion 'time period', it is only natural to say that attempts at periodization are never neat: one period flows into another or the same name may be used for periods that occur at different times in different regions, for instance, "εμφύλιος" (civil war) is a time period that determines different times from the ancient history of Greece to our days.

However, time period names are used widely in both the everyday language and the scientific jargon and are very useful in documentation. Furthermore, their study could reveal how human communities, here the Greek speaking communities, coin period names, for example which events lend their names and which linguistic means are employed to coin a time period name.

To the best of our knowledge, there is no structured collection of time period names available about Modern Greek History; indicatively, in the relevant lemma in the Wikipedia such information is extremely lean. Furthermore, there is no study concerning the linguistic characteristics of these terms. So, the first step is to develop a corpus of time period names. The second step is their organization into ontologies and timelines. In this paper we report on the first step.

We report on the development of vocabularies of historical period names from Web acquired specific corpora. Web data would cater for the sparcity of Greek corpora dense in time period names and for the dynamic nature of these terms. We focused on period names of the 19th and 20th century of Greek History. For data collection from the Web we used the ILSP - Focused Monolingual Crawler (FMC) (Mastropavlos, et al., 2011).

This paper is organized as follows: Section 2 briefly presents related work. Section 3 describes the methodology we developed to collect Web corpora and identify the time period names. Section 4 introduces some observations concerning the strategy Greek communities use to coin new time period names. Conclusions are given in Section 5 and plans for future work in Section 6.

2. RELATED WORK

Time period names have only recently attracted the ICT research community because Semantic Web has increased needs in standardized documentation and linked data. Structured collections, for instance ontologies, of time period names are sparse however (Berman, 2011).

Feinberg et al., (2003) pioneered work on time period names. They treat them as gaz-

etteers and present a Content Schema that specifies the information required to define a period name. Furthermore, drawing on data from the University of California MELVYL catalog and several timelines available on the Web, they discuss the issue of distinguishing between time period names and event names. Although duration is a semantic criterion that favors the time period reading of a term, it is not always conclusive. For instance, they notice that Kennedy assassination is used as a micro-era period name despite the fact that it clearly refers to an instantaneous event. The authors propose that a set of criteria are used including the role of the name in context and duration of the time interval denoted.

Petras et al. (2006) developed a prototype Time Period Directory with time period names and events extracted from the Library of Congress subject headings. Drawing on Feinberg et al., (2003), they too argue that a time period directory could work as a place name gazetteer does because (1) as locations are referred to by place names similarly spans of time are commonly referred to by period names, such as "Napoleonic wars" and (2) time periods have a geographical aspect as gazetteer entries have a period aspect (Buckland, et al., 2004). They developed a Time Period Directory with 2,000 entries derived from the Library of Congress Subject Headings. Drawing on the Feinberg et al. (2003) Content Schema, they proposed a Content Standard and the relevant XML Schema that has been adopted by the ECAI (http://ecai.org/).

The DDBC Time Authority Database (http://authority.ddbc.edu.tw/docs/open_content/) is one of a group of Authority Databases provided by Dharma Drum Buddhist College (DDBC). It contains detailed Chinese calendar data from the beginning of the Qin dynasty to the current day. The major purpose of DDBC is to provide complete Chinese calendar information that can be used by external services and applications. Time period names play an important role in the organisation of the DDBC data.

Time period names have been considered as crucial parts of next generation gazetteers (Berman, 2011) that will link together Place Names (Place Name Authorities, Historical GIS and geonames) with Chronologies (Administration Periods, Timelines of events and Time Period Names Index) and Entity Definitions.

3. METHODOLOGY FOR COLLECTING (GREEK) TIME PERIOD NAMES

We report on the development of collections of Greek time period names. This work consists of two parts: (1) collection of Web corpora dense in time period names (2) multidimensional annotation of terms. Our work differs from existing work on period names in that there were no structured resources, such as the Library of Congress archives or substantial timelines, to refer to. Instead, we had to resort to fully unstructured resources and in fact, discover ways of collecting the right ones. Furthermore, given that the contexts where the terms occurred were of a general nature, we had to disambiguate among more term readings than simply those of a period name and an event.

We confined ourselves to the 19th and the 20th century of Greek history (1821: the successful Greek Independence Revolution against the Turkish Occupation) in order to simplify things. To develop a rich collection of period names from the Greek History of the 19th and the 20th century, a specific domain corpus with different types of text should be used. The need for a large variety of texts is intensified by the fact that period names should be treated as a dynamic phenomenon.

The available Greek corpora are of small or medium size and certainly not dense in such terms. The existing general purpose Greek corpora, Hellenic National Corpus being the standard example (Gavrilidou, 2002), as well as Corpus Greek Texts (Goutsos, 2010) have not been built for use in specialized domains, such as history, medicine or education.

Fortunately, tools for easily developing corpora rich in the respective material of interest are available nowadays. We used a revised version of the Focused Monolingual Crawler (FMC) (Mastropavlos, et al. 2011) developed by the Institute for Language and Speech Processing (ILSP/ "Athena" RIC). The collected corpora were processed manually to extract candidate time period names and their contexts. Then, a battery of linguistic criteria was used to distinguish between names that denoted time periods and names that denoted something else, most often events.

3.1 Focused Monolingual Crawler (FMC)

FMC is a system that explores the Web and downloads pages with text related to a specific domain. Also, it includes components for all the tasks required to acquire domain-specific corpora from the Web. Due to its modular architecture, each of these components can be easily substituted with alternatives of the same functionality. Furthermore, the system is available as an open-source Java project (http://nlp.ilsp.gr/redmine/projects/ilsp-fc/). Comprehensive documentation on how to get setup and run it, is available (http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki).

Given a narrow domain defined by a certain topic and a language, FMC is fed with two input datasets: (i) a list of topic definition (multi-)word terms and (ii) a list of topic related URLs (Skadina et al., 2012). The user can configure FMC in a variety of ways, for instance set file types to download, domain filtering options, self-terminating conditions, crawling politeness parameters, etc".

The crawler first visits the web pages that are initially provided by the user and fetches related ones. Next, it classifies fetched pages as relevant to the targeted domain and extracts links from fetched web pages. Finally, it adds them to the list of pages to be visited and repeats this cycle. A typical workflow for acquiring monolingual domain-specific data is illustrated in Figure 1.

3.2 Experiments with the FMC

Three experiments were conducted with FMC. Greek was used for all the three experiments. Greek is not an under-resourced language, as the basic information technology is available and has a relatively substantial presence in the Web (Berment, 2004). On the other hand, there is a clear need for large Greek corpora and electronic lexica and this fact underlines the importance of FMC for Greek.

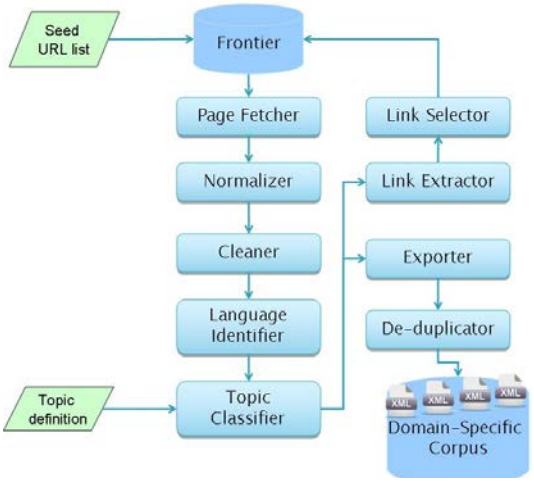

Figure 1 A typical workflow of FMC.

In all the three experiments the input to the FMC consisted of a seed URL list that initialized the crawler's frontier. In our case, a possible initial web page would be the Wikipedia page for "Τουρκοκρατία" (Turkish Occupation). As explained in Section 3.1, the input also included the domain definition that consisted of terms describing the domain.

We set off with a TermList of twenty five terms. Figure 2 shows a subset of these terms. The terms were collected from the High School History textbooks. Our aim was to create a first sample, a list just to feed the crawler. In this first experiment only a dozen of URLs retrieved from search engines were given as an input. Mostly, they were Web pages of museums, libraries and wikis.

In all the three experiments, we set the time parameter of the crawler to a single hour of work. We arrived to this decision because we wanted the size of our corpus to be manageable for a manual retrieval of period names.

TermList_Exp1
Ελληνική επανάσταση-Greek revolution
20ος αιώνας-20th century
Ανατολικό Ζήτημα-Eastern Question Περίοδος ανεξαρτησίας-Independence period
Περίοδος αντιβασιλείας-Regency period
Α' Βαλκανικοί Πόλεμοι-1st Balcan Wars
Ελληνοϊταλικός Πόλεμος-Greek-Italian War
Μικρασιατική Καταστροφή Asia Minor Destruction
Εμφύλιος Πόλεμος-Civil War
Κατοχή-German Occupation
Δικτατορία του Παπαδόπουλου Papadopoulos' Dictatorship
Μεταπολίτευση (period immediately after the 1967-1973 Junta)

Figure 2 TermList – Experiment 1.

The output of the first experiment contained almost 500 URLs. With an advanced search we found out that the result would be better if our TermList was enriched with more period names.

Apart from period names, in the subsequent two experiments we used key-words that were not historical periods, such as the name of the protagonist of a fight, a war or a military movement as well as names of locations.

Moreover, we added words that were found in the contexts where historical names occurred. The aim was to enhance our searching tools and obtain more results relevant to the periods in question. In order to identify such words, we studied the contexts where the term "Τουρκοκρατία" (Turkish Occupation) occurred. Since the period is dominant in Modern Greek history, we knew that there would be a big amount of documentation for it. In that way we identified several syntactic environments and enriched our TermList for the experiments that followed.

TermList_Exp2
Τουρκοκρατία-Turkish Occupation
Οθωνική περίοδος-Otto period
Β' Παγκόσμιος Πόλεμος-2nd World War
Δικτατορία του Μεταξά Metaxas' Dictatorship
Περίοδος της Αντίστασης-Resistance Period
Στα χρόνια
Κατά την περίοδο
Την εποχή
Τα γεγονότα της
Κατά τη διάρκεια **Επί πρωθυπουργίας**

Figure 3 TermList – Experiment 2.

Last, we noticed that certain words functioned as heads of period names such as the word "πόλεμος" (war) or "εποχή" (era) or even "περίοδος" (period). Those words were also included in our term list.

Furthermore, in the next two experiments we enriched the URLs list with web pages that would use more colloquial language and would capture the dynamic nature of period names, such as newspapers, blogs and forums.

Fig. 3 shows the results of the three experiments:

Parameters	Ex1	Ex2	Ex3
TermList	25	75	180
Language	Gr	Gr	Gr
URLlist	12	25	50
MaxTime	1h	1h	1h
	Results		
XML/URL	487	911	1622

Figure 4 The input and the output of the three experiments.

3.3 Annotation

We annotated our corpus manually. We retained a 50-word context before and after

the terms of interest. Apart from this information, terms were annotated for the following features: syntactic structure (for instance Adjective+Noun), syntactic role in the text (subject, object etc), allomorphs of the term in the collected corpora, close context that helped disambiguating the reading of the term between that of a time period and that of an event, whether the term denoted a fact or a time period, whether the term denoted a specified time period or part of a specified time period, persons, places and other period names related with the term, the actual dates related with the term and the URL where the term was retrieved from.

We used contextual information to disambiguate between a time period reading and other readings, the event reading being the dominant but not the only alternative used. As a working example we will use the term Εθνική Αντίσταση (National Resistance) that refers to the Greek resistance during the German Occupation (1941-1944). The term returned 25 examples, of which 8 were classified as period names.

We first set apart metonymic usages such as 'agent' (1) and 'moral imperative' (2).

(1) Η Εθνική Αντίσταση κέρδισε την πρώτη μάχη της σε ανοιχτή αναμέτρηση με τις δυνάμεις της φασιστικής βίας.

'National Resistance wined its first battle in an open struggle against the powers of fascist violence'.

(2) Αντίθετα, μετά την Εθνική Αντίσταση του ΕΑΜ, το πρόταγμα για Λαοκρατία κρατά συμπαγές το ΚΚΕ...

'On the contrary, after the National Resistance, the ethical imperative Rule-Of–the-People pulls the Greek Communist Party together' …

Next, we classified as event names terms that occurred in the following contexts:

–When the term was used to explain the word 'event'

-Contexts where a set of actions were denoted. For instance, in (3), contrary to an event, a time period just occurs and can not be 'organised'. A similar context is 'X took an active role in <TERM>'

- Lists of events that included the term in question.

(3) Στην αρχή του πολέμου ο κόσμος έμεινε άφωνος... αλλά στη συνέχεια οργανώθηκε η Εθνική Αντίσταση

'When the war started, people were left speechless… but then the National Resistance was organized'.

We classified as time period names terms that occurred in the contexts listed below:

- Titles (ambiguous between an event and a time period reading)
- The term offered the time contour for the event that was described (4)
- Contexts where a 'set of coherent …cultural manifestations' reading (Crofts et al., 2004) was allowed (5).

(4) Μετά από όλες τις ταλαιπωρίες που πέρασε πολεμώντας στην Μικρά Ασία,..., κατόπιν στο Αλβανικό και αργότερα στην Εθνική Αντίσταση όπου ήταν Διοικητής στο 52 Σύνταγμα του ΕΛΑΣ στο Λιανοκλάδι...

'After all the suffering he went through fighting in Asia Minor, …, then in the Albanian War and later during the National Resistance when he was Commander of the 52th Regiment of ELAS at Lianocladi…'

(5) …διοργανώνει, σε συνεργασία με το Δήμο Ηλιούπολης, εκδήλωση με θέμα: «Η Ελληνίδα Γυναίκα στην Εθνική Αντίσταση και η σύγχρονη κοινωνική πραγματικότητα.

'…organizes, in cooperation with the Ilioupoli Municipality, a discussion with the subject "The Greek Woman at the National Resistance and the modern social situation'.

As is the rule, term sense disambiguation is not a clear-cut job. We plan to include some inter-annotation experiments in our future work.

We started our study with only 19 period names and no contexts of usage. We now have a list of 78 period names, each one exemplified with an average number of 16 contexts of usage. Given that we achieved these results with only 3 hours of Web corpora collection with FMC, we conclude that our method can be used

successfully to develop corpora dense in domain specific information that evolves dynamically. Certainly, identification of time period names in the accumulated corpora is a laborious task that is performed manually given the present state-of-the-art in NLP in terms of recall and precision. However, the collection of appropriate corpora is actually the hardest and least attractive part of the work. Our method reduces to a minimum the effort for appropriate corpora collection. In addition, the annotated corpora that have been developed at the course of this work form a recourse that could be used as a training/test set for developing a module for automatic identification of period names.

4. CONSTRAINTS ON THE CREATION OF TIME PERIOD NAMES

A first picture of the semantic constraints that control the process of coining time period names has emerged from our work. And in fact, as native speakers of Greek, we were surprised to find out that the Greek communities do not coin period names out of names of treaties, military movements and battles. Figure 4 offers an idea of the overall picture.

Period Names	Events
Καποδιστριακή Περίοδος The period of Capodistrias	Κίνημα στο Γουδί (military) Movement at Goudi
Παλινόρθωση When the Greek King resumed his reign	Συνθήκη των Σεβρών Treaty of Sevres
Χρόνια της Κρητικής Πολιτείας The years of the Cretan State	Μικρασιατική Καταστροφή Asia Minor Destruction
Εθνικός Διχασμός National Division	Κίνημα του Ναυτικού Movement of the Navy
Δεκεμβριανά The events of December	Πραξικόπημα Coup
Εμφύλιος Civil (War)	Μοναστηριακό ζήτημα Monastery Issue
Είσοδος της Ελλάδας στην Ε.Ε. Admission of Greece to the EU	Κρητική Επανάσταση Cretan Revolution

Figure 5 Period Names – Facts/Other Readings.

The explanation that treaty signing, military movements and battles are events of a rather short duration may be put forward for this tendency of Greek. However, strong counterexamples to this explanation exist. For instance, we were very surprised to find out that "Μικρασιατική Καταστροφή" (Asia Minor Destruction) that describes an extended period (at least the biggest part of the year 1922) was never used with a time period reading in our data.

5. CONCLUSIONS

We have described a method for collecting a substantial number of period names from unstructured Web data. The method involves techniques for collecting specific domain corpora dense in the desired terms by using the Focused Monolingual Crawler of ILSP. The method also comprises the identification of linguistic contexts where period names occur. The method is useful to languages that have a presence in the Web but have no corpora appropriate for time period name retrieval. In addition, the method may be useful to languages with richer resources because time period names are a dynamic phenomenon that can be better described with corpora that reflect the current usage of language.

6. FUTURE WORK

In the immediate future we plan to take the following steps:

-Verify our sense disambiguation results with targeted inter-annotation experiments

-Use the Content Standard developed by Petras et al. (2006) and perhaps extend it with necessary information of a linguistic

nature in order to encode as gazetteers the time period names we have collected

-Develop an ontology of Greek period names that will be useful to ICT applications and lexicography

- Enhance/ modify existing modules (e.g. TimeEL proposed by Prokopidis et al 2009) for automatic recognition of temporal expressions in Greek texts

-Develop timelines for the Wikipedia and for educational purposes (visualization of data). As Greece is a country consisting of very many small places, each one with its own varied and turbulent history, such timelines will be of immense help not only to students and teachers but to the layman who is interested in history—let alone the researchers.[1]

[1] An excellent example of the educational usages of such timelines is the Greek term 'Τουρκοκρατία' (Turkish Occupation). Here, we describe the complications of the case of only one Greek town, namely the town of Naplion that is situated in Eastern-Central Peloponnese. There are two periods with the same name for this town. The first one was between 1540 and 1687 and the second one was from 1715 to 1822. The situation gets very complex if all the areas where the term applies (they are many more than the ones included in the today Greek territory) are considered as there are areas that were under the Turks for 16 years only and areas that were under the Ottoman regime for five centuries.

REFERENCES

Berman, M. (2011). *Extending Gazetteers with Time and Entity Relationships*. Historical Gazetteer Elements: Temporal Frameworks. Track on Historical Gazetteers
Part of the Symposium on Space-Time Integration in Geography and GIScience
co-sponsored by Harvard Univeristy's Center for Geographic Analysis and the AAG, Wednesday-Friday, April 13-15, AAG 2011, Seattle, WA

Buckland, M. and Lancaster, L. (2004) *Combining Time, Place, and Topic: The Electronic Cultural Atlas Initiative*, D–Lib Magazine, volume 10, number 5 (May), at http://www.dlib.org/dlib/may04/buckland/05buckland.html, accessed 2 June 2006.

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. (2004) *Definition of the CIDOC Conceptual Reference Model* (version 4.0).

DDBC Time Authority Database (http://authority.ddbc.edu.tw/docs/open_content/)

Doerr, M., Kritsotaki, A. and Stead, St. (2003) *Thesauri of Historical Periods – A Proposal for Standardization*, (http://www.cidoc-crm.org/).

Feinberg, M., Mostern, R., Stone, S. and Buckland, M. (2003) *Application of Geographical Gazetteer Standards to Named Time Periods.* Technical Report, Electronic Cultural Atlas Initiative, Berkeley.

Gavrilidou, M., (2002) *The Hellenic National Corpus on-line*, Revue Belge de Philologie et Historie 80, pp. 1003-1015

Goutsos, D., (2010) *The Corpus of Greek Texts: a reference corpus for Modern Greek*, Corpora. Vol 5, pp. 29-44.

Harvard University. Chinese Historical GIS Project. Available at <http://www.fas.harvard.edu/~chgis/>.

ISO/CD 21127 (2002) *Information and documentation – A reference ontology for the interchange of cultural heritage information*

Mastropavlos, N. and Papavassiliou, V. (2011) *Automatic Acquisition of Bilingual Language Resources.* In Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece.

Petras, V., Meiske, M., Larson, R., Zernecke, J., Carl, K. and Buckland, M. (2005) *Leveraging Library of Congress Subject Headings to improve Search for Events – A Time Period Directory*.

Petras, V., Larson, R. and Buckland, M. (2006) *Time Period Directories: a Metadata Infrastructure for Placing Events in Temporal and Geographic Context.* Joint Conference on Digital Libraries, Chapel Hill, NC, USA.

Prokopidis, P., Desipri, E., Papageorgiou, H. and Markopoulos, G. (2009) *TimeEL: Recognition of Temporal Expressions in Greek texts*. In Proceedings of the 9th International Conference of Greek Linguistics, Chicago, Illinois, USA.

Skadiņa, I., Aker, A., Μαστροπαύλος, N., Su, F., Tufis, D., Mateja, V. et al. (2012). *Collecting and Using Comparable Corpora for Statistical Machine Translation.* In On-Line Proceedings of the LREC2012 Conference on Language Resources and Evaluation, pages 438-445. Istanbul, Turkey.

Support for the Learner: Time Periods, at http://ecai.org/imls2004/timeperiods.html, accessed 2 June 2006.
Wikipedia. List of Themed Timelines. 2004. Available at <http://en.wikipedia.org/wiki/List_of_themed_timelines>.