20. Creation of an electronic dictionary of sports terminology

**Kyriaki Ioannidou, Anthie Kyriakopoulou,
Olympia Tsaknaki, Rania Voskaki**

SUMMARY

This paper is about the creation of an electronic dictionary of sports terminology, which will be used by computers in natural language processing systems. This dictionary is based on the Greek version of a multilingual sports terminology database, which was processed within the framework of the Euradic project as part of the Technolangues project. The terms were first translated into Greek. The next step was to transfer the database entries to the electronic terminology dictionary including verbs, simple and compound nouns, adjectives, simple and compound adverbs, initialisms, frozen expressions. For the creation of this dictionary, the computational linguistics team of the Natural Language Processing Unit, Aristotle University of Thessaloniki, and Gaspard Monge Institute, University of Marne-la-Vallée used the same codification and automatic inflection programmes that were used for the creation of their general electronic dictionaries. The research was conducted by the said team, whose studies aim at a detailed and formalised description of Modern Greek, the final objective being the recognition of linguistic data by natural language processing systems.

ΔΗΜΙΟΥΡΓΙΑ ΗΛΕΚΤΡΟΝΙΚΟΥ ΛΕΞΙΚΟΥ ΑΘΛΗΤΙΚΗΣ ΟΡΟΛΟΓΙΑΣ

**Ράνια Βοσκάκη, Κυριακή Ιωαννίδου,
Ανθή Κυριακοπούλου, Ολυμπία Τσακνάκη**

ΠΕΡΙΛΗΨΗ

Η παρούσα μελέτη αφορά τη δημιουργία ενός ηλεκτρονικού λεξικού αθλητικής ορολογίας για χρήση σε συστήματα αυτόματης ανάλυσης φυσικών γλωσσών. Το λεξικό αυτό βασίζεται στην ελληνική έκδοση μίας πολύγλωσσης βάσης δεδομένων αθλητικής ορολογίας, η επεξεργασία της οποίας έγινε στα πλαίσια του προγράμματος Euradic, που εντάσσεται στο πρόγραμμα Technolangues. Αρχικά, μεταφράστηκαν οι όροι στα ελληνικά. Στη συνέχεια έγινε η μετάβασή τους από τη βάση δεδομένων στο ηλεκτρονικό ορολογικό λεξικό το οποίο συμπεριλαμβάνει ρήματα, απλά και σύνθετα ονόματα, επίθετα, απλά και σύνθετα επιρρήματα, αρκτικόλεξα, παγιωμένες εκφράσεις. Για τη δημιουργία του λεξικού αυτού χρησιμοποιήθηκαν τα προγράμματα κωδικοποίησης και αυτόματης κλίσης που έχουν χρησιμοποιηθεί για την κατασκευή των γενικών ηλεκτρονικών λεξικών της Νέας Ελληνικής από την ομάδα υπολογιστικής γλωσσολογίας της Μονάδας Αυτόματης Επεξεργασίας Φυσικών Γλωσσών του Α.Π.Θ. και του Institut Gaspard Monge του Πανεπιστημίου της Marne-la-Vallée. Η έρευνα πραγματοποιήθηκε από την προαναφερθείσα ομάδα, οι μελέτες της οποίας αποσκοπούν στη λεπτομερή και τυποποιημένη περιγραφή της Νέας Ελληνικής με σκοπό την αναγνώριση των γλωσσικών δεδομένων από συστήματα αυτόματης ανάλυσης φυσικών γλωσσών.

0 PRESENTATION

The starting point of this study was the Euradic project as part of the Technolangues project. It was financed by the French Ministry of Economy, Finance and Industry. The objective was the creation of a multilingual sports terminology database in the following languages: French, English, Arab, German and Greek. The partners were the Publishing House *La Maison du Dictionnaire* (France), University of Rennes II (France), Aristotle University of Thessaloniki (Greece), University of Tunis (Tunisia), Centrum für Informations- und Sprachverarbeitung (Germany). The database is in Access format and contains 37 000 terms of summer and winter Olympic sports. The first part of the database (summer Olympic sports) was made available gratis during the Athens 2004 Olympic Games (http://www.at-lci.com/euradic/). The full version will be soon made available on Internet for a fee.

Initially the Greek terms were translated from French. The next step was to generate the inflected forms of the Greek terms in order to use them in NLP systems. For this purpose we used the GenereFlexion programme. The objective of this effort was to create an electronic dictionary - as complete as possible - of sports terms that would be applied in sports electronic corpora. This programme is a semi-automatic procedure combining the canonical forms of entries with the respective suffixes.

1 EURADIC PROJECT: THE GREEK VERSION

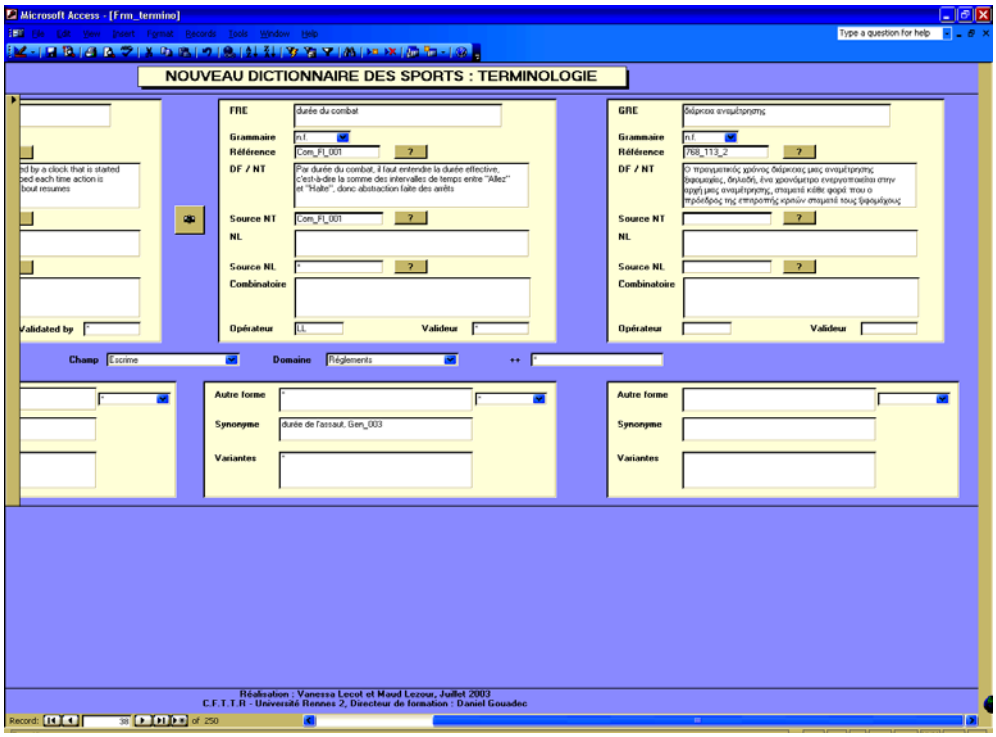
1.1 STRUCTURE OF DATABASE AND TERMINOLOGY CARDS

The Greek version of the database was based on the French and English ones, which the University of Rennes II made available to the rest of the partners. The French and English terms were classified into 73 sport fields for the summer Olympic sports (such as rowing, badminton, baseball, basketball, boxing, canoe-kayak, etc.) and 14 for the winter ones (such as alpine ski, snowboarding, skeleton, etc.) The terms were also classified into the following domains: arbitrage, training, event, equipment, general, person, physiology/anatomy, regulations, technique, apparel, field.

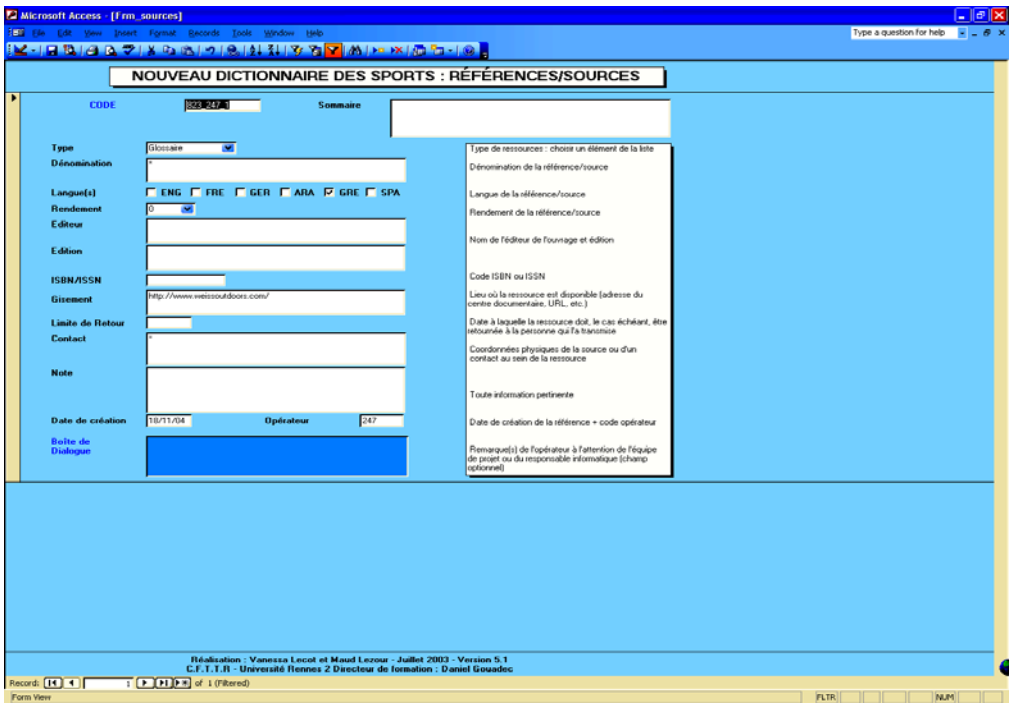
A terminology card was created for each term and in all five languages. The terminology card contained the following information: the term itself, grammatical category, reference, author's name, the respective field and domain mentioned above and synonyms.

First, all sport fields and domains were translated, because each Greek term had to be classified under one of them. As for the Greek terminology cards, the same general structure was kept. Necessary adjustments were made so as to adapt the card structure to the Greek data: the Greek grammatical categories are different from the French and English ones. The grammatical classification of the Greek terms included interjections,

adverbs, adjectives, masculine nouns, feminine nouns, neutral nouns, masculine nouns in plural, feminine nouns in plural, neutral nouns in plural and verbs (see below an example of a terminology card). At this stage no distinction was made between simple and compound words. This classification was modified later on, when the data were further processed to be integrated in the Greek electronic dictionaries.



A separate card was used for the sources. This card provided the following information: term code, type of used source, denomination, language, web address and date of creation. Here is an example of the source cards:



1.2 TERM TRANSLATION

The translation of the terms was proved to be complicated, due to the significant lack of reliable terminology tools and multilingual sports dictionaries comprising a Greek section. Dictionaries (monolingual, bilingual and multilingual, general and specialised) and glossaries, regulatory texts, manuals, other terminology databases or banks were used as sources for the translation of the terms. Internet was widely used to get to these sources. Finding the regulatory texts was not always an easy task, as some sports are not very popular, especially in Greece (e.g. Nordic combined ski, bobsleigh, curling*)*. Translation of winter sports terms proved also to be a difficult task, mainly because most of these sports are not even practised in Greece (or are practised at a very low scale).

At the translation stage, we found many terms with established translation equivalent(s) in Modern Greek (e.g. *ποδόσφαιρο/football)*. Among these terms there are also translation loans (*ημιτελικός/semi-final),* semantic loans (*Ολυμπιακοί Αγώνες/Olympic Games),* reborrowings (*κάρτα/card),* loans adapted to the inflectional system of Modern Greek with the addition of prefixes and/or suffixes (*επινικελωμένο/nickel plated)* and initialisms (*FIG-Διεθνής Ομοσπονδία Γυμναστικής/FIG*). The translation of the names of sports federations or organisations required extended research, as the official translation had to be found.

However, we found terms with no translation equivalent(s) in Modern Greek. These terms were either composed of one word (e.g. *bumper/αναστολέας*) or of more than one word (e.g. *εμπρόσθιος αναστολέας στερέωσης/front fastening brackets, έγγραφο για αθλητή που αποσύρεται λόγω σοβαρού τραυματισμού/"athlete withdrawn due to serious injury" document*). In these cases we were obliged to create new terms by respecting the language structure. We must notice that a detailed analysis of the created neologisms is not possible in the framework of this study.

2 TRANSFERRING THE DATA INTO THE GREEK ELECTRONIC DICTIONARY

2.1 ELECTRONIC MORPHOLOGICAL DICTIONARY OF MODERN GREEK

This study is part of the joint activities of the laboratories of computational linguistics of the Aristotle University of Thessaloniki[1] and of the University of Marne-la-Vallée[2], aiming to developing a complete and formalised description of the Modern Greek language.

Formalised electronic dictionaries are needed for the natural language processing. By the term *electronic dictionaries* we mean the dictionaries created to be used by computers in natural language processing systems. The information contained in the electronic dictionaries should be as detailed and complete as possible. These dictionaries are constantly updated with new entries and their respective linguistic (morphological, syntactical and semantic) properties and grammars. Therefore, electronic dictionaries should include information about all inflected types of verbs, simple and compound nouns and adjectives. The *simple words* are sequences of letters comprised between two consecutive separators. The *compound words are* sequences including at least two simple words and at least one separator [1], [6]. In Greek there are the following separators [5]: the space *(Ολυμπιακοί Αγώνες/Olympic Games)*, the hyphen *(διαιτητής-κριτής/referee)* and the apostrophe and space *(κατ' ευθεία πάσα/volley pass)*.

The creation of the Greek electronic dictionaries was based on the DELA system, which was developed in LADL[3] under the supervision of Maurice Gross. The DELA system includes DELAS, which is a dictionary of simple words, and DELAC, which is a dictionary of compound words. From DELAS and DELAC all inflected typed are automatically generated. They are included in DELASF and DELACF respectively. These dictionaries provide only

[1] http://linginfo.frl.auth.gr/.
[2] http://ladl.univ-mlv.fr/.
[3] Laboratoire d'Automatique Documentaire et Linguistique.

morphological information[4].

Today, the dictionary of canonical forms of Modern Greek contains[5]:

- 70 000 simple words;
- 18 000 verbs;
- 40 000 adjectives;
- 16 000 simples and compounds adverbs;
- 28 000 compound words;
- 50 000 proper nouns;
- 2 000 country names;
- 1 000 simple and compound grammatical words.

To these entries there were added 940 simple words and 18 000 compound words of summer and winter Olympic sports terminology.

2.2 RESTRUCTURING THE DATA

At this stage the data were reclassified in a formalised way. Significant modifications were made to the initial classification. A distinction was made between simple and compound words. For instance, *καλαθοσφαίριση/basketball* was classified as a simple noun, whereas *γραμμή οριοθέτησης/boundary line* was classified as a compound noun. Both terms had initially been classified simply as nouns in the terminology database. The compound words are further divided in different categories, such as:

- A N, standing for the *Adjective Noun* type of compound words: *αγωνιστικός (A) χώρος (N) (wrestling area);*
- N N_{gen} standing for the *Noun Noun in genitive* type of compound words: *γραμμή (N) τέρματος (N) (goal line);*
- N DET N_{gen} standing for *Noun Determiner Noun in genitive* type of compound words: *σύνθεση (N) της (DET) ομάδας (N);*

At the initial stage many terms had been classified as verbs, simply because the entry started with a verb. In many of these cases the new classification was different because the said verb was just the support verb. The element containing the semantic weight was the predicative noun [2], which means that the noun and not the verb should constitute an entry of the electronic dictionaries: *κάνω βήματα/to travel, κάνω τάκλιν/to tackle.* In this case, a full syntactic and semantic analysis will be obtained by the creation of a lexicon-grammar table

[4] Syntactic and semantic information is provided in the lexicon-grammar tables; lexicon-grammar is a syntactic-semantic electronic dictionary.

[3]. The formalism of lexicon-grammar tables allows the detailed syntactico-semantic description of predicative nouns such as *τάκλιν/tackle.* This means that all the predicative nouns should be examined in relation with the support verbs which accompany them.

In addition, at the translation stage, an acronym was given as a synonym for a term with no additional information: *ΙΑΑF/I.A.A.F./Διεθνής Ένωση Ομοσπονδιών Στίβου/International Association of Athletics Federations.* At the morphological description stage the distinction was made clear and acronyms were classified as such.

2.3 GENEREFLEXION

GenereFlexion was used for the automatic inflection of all entries of the general electronic dictionaries and therefore for the automatic inflection of all sports terms. GenereFlexion is an automatic inflection programme, created by T. Kyriacopoulou and S. Mrabti [5]. It was developed in C, it launches in DOS or in UNIX and only plain text files (*.txt) can be processed with it. For the automatic generation of all inflected forms the programme uses three files. The first file contains the canonical forms together with a symbol to indicate the part of speech they belong to (*N* for nouns, *A* for adjectives, *DET* for articles, *ADV* for adverbs, *CONJ* for conjunctions and *PREP* for prepositions), the respective morphological code and the symbol *“* in case the stress moves when the word is inflected. Specific filters can be used, e.g. *S* when the entry is used only in singular and *P* when the entry is used only in plural. In the case of compound words one symbol indicating the part of speech is used after each word of the compound lexical unit and a second one is used at the end of the sequence after the symbol used for the last word of the compound lexical unit, to indicate the part of speech of the compound word as a whole. In certain compound lexical units one of the consisting words is used only in a specific case: a specific filter is used to indicate the case (*N* for nominative, *G* for genitive, *A* for accusative and *V* for vocative) and, if necessary, the number of the word.

ισοβαθμία.N232,N (tie on points)

ήττα.N232 στα.PREP σημεία.N311,N,-AP3 (loss on points)

σώμα".N363 διαιτησίας.N232,N,-GS2 (officiating team)

ορεινή.A10 ποδηλασία.N232,N,S (mountain bike)

ανοιχτή.A10 πίστα.N232,N (outdoor track)

[5] 30% of this data are available on the Internet.

The second file contains the inflectional vectors:

N232.2.α,ας,α,α,ες,°ών,ες,ες

N311.3.ο,ου,ο,ο,α,ων,α,α

N363.3.0,τος,0,0,τα,"΄"°των,τα,τα

A10.2.ή,ής,ή,ή,ές,ών,ές,ές

In the third file all inflected forms are generated automatically.

ισοβαθμία,.N:Nfs:Afs:Vfs

ισοβαθμίας,ισοβαθμία.N:Gfs

ισοβαθμίες,ισοβαθμία.N:Nfp:Afp:Vfp

ισοβαθμιών,ισοβαθμία.N:Gfp

ήττα στα σημεία,.N:Nfs:Afs:Vfs

ήττας στα σημεία,ήττα στα σημεία.N:Gfs

ήττες στα σημεία,ήττα στα σημεία.N:Nfp:Afp:Vfp

ηττών στα σημεία,ήττα στα σημεία.N:Gfp

σώμα διαιτησίας,.N:Nns:Ans:Vns

σώματος διαιτησίας,σώμα διαιτησίας.N:Gns

σώματα διαιτησίας,σώμα διαιτησίας.N:Nnp:Anp:Vnp

σωμάτων διαιτησίας,σώμα διαιτησίας.N:Gnp

ορεινή ποδηλασία,.N:Nfs:Afs:Vfs

ορεινής ποδηλασίας,ορεινή ποδηλασία.N:Gfs

ανοιχτή πίστα,.N:Nfs:Afs:Vfs

ανοιχτής πίστας,ανοιχτή πίστα.N:Gfs

ανοιχτές πίστες,ανοιχτή πίστα.N:Nfp:Afp:Vfp

ανοιχτών πιστών,ανοιχτή πίστα.N:Gfp

3 CONCLUSIONS AND PERSPECTIVES

This study consisted in creating a multilingual sports terminology database, the Olympic Games being the reference point. This database was further processed to be integrated in electronic dictionaries. It was the first organised effort to include specialised terminology in our electronic dictionaries. Other specialities will follow to complete our dictionaries with technical terms. The second part of the research is placed in the framework of morphology. This means that further systematic description of the terms is needed at syntactical and semantic levels. More specifically, our objective is to create local grammars and lexicon-grammar tables [3], [7], in order to achieve the most complete possible recognition of electronic sport texts at all levels of analysis. The local grammars will concern the representation of common terms by finite state automata at the paradigmatic axis (e.g. football/basketball/volleyball ball). The lexicon-grammar tables will describe the predicative nouns and respective support verbs, providing all the necessary syntactical and semantic information.

References

[1] Αναστασιάδη-Συμεωνίδη, Α. *Η Νεολογία στην Κοινή Νεοελληνική*: Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής του Α.Π.Θ., Παράρτημα 65, Θεσσαλονίκη, 1986.

[2] Giry-Schneider, J. *Les prédicats nominaux en français. Les phrases simples à verbe support.* Droz, Genève 1987.

[3] Gross, M. *Méthodes en syntaxe – Régime des constructions complétives.* Hermann, Paris, 1975.

[4] Κυριακοπούλου, Α. "Συγκριτική μελέτη του ειδικού λεξιλογίου του ποδοσφαίρου στη Νέα Ελληνική και τη Γαλλική και μέθοδοι αναπαράστασής του". *Ελληνική Γλώσσα και Ορολογία, Ανακοινώσεις 4ου Συνεδρίου*. Αθήνα: Τεχνικό Επιμελητήριο Ελλάδας, 2003, σσ.: 290-299.

[5] Kyriacopoulou, T. - Mrabti, S. - Yannacopoulou, A. «Le dictionnaire électronique des noms composés en grec moderne», *Lingvisticæ Investigationes* 25:1, Amsterdam/Philadelphia, John Benjamins, 2002, pp. 7-28.

[6] Silberztein, M. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris : Masson, 1993.

[7] Sklavounou, E. *Λεξικό-γραμματική σύνθετων ονομάτων. Υποβοηθητικά ρήματα (εφαρμογή στο ειδικό λεξιλόγιο του τένις για την Ελληνική, Γαλλική και Αγγλική)*

Αναπαράσταση με πεπερασμένα αυτόματα, μεταπτυχιακή εργασία, Φιλοσοφική Σχολή, Α.Π.Θ., Θεσσαλονίκη, 1994.

Kyriaki Ioannidou

Postgraduate student of the Interdisciplinary Postgraduate Studies Programme in Sciences and Technologies of Language and Communication of the Aristotle University of Thessaloniki

Address: Α.Π.Θ. Γαλλικό Τμήμα, GR-54 124, Tel.: (+30) 2310 99 75 16.

Anthie Kyriakopoulou

Phd student, Institut Gaspard Monge, Université de Marne-la-Vallée

Olympia Tsaknaki

Post-doc researcher, Institut Gaspard Monge, Université de Marne-la-Vallée

Rania Voskaki

Phd student, Institut Gaspard Monge, Université de Marne-la-Vallée

Laboratoire d'Informatique
Equipe d'Informatique linguistique
5 Bd Descartes, Champs-sur-Marne
7454 Marne-la-Vallée Cedex 2

Tél. : (+33) (0)1 60 95 77 15