Multi-TALE

The generation of MULTI - lingual specialised lexicons by using augmented TAgger - LEmmatizers

Commission of the European Communities
Multilingual Action Plan Programme
MLAP contract 93/04

Project Summary

The aim of MULTI-TALE is to develop a syntactic-semantic tagger-lemmatizer that in a multi-lingual environment assists lexicologists in the field of health care informatics to classify automatically specialised terms in a standardisation perspective of the medical sublanguage. MULTI-TALE will therefore investigate how two existing grammatical tagger-lemmatizers, developed for general (i.e. domain-independent) language, are suitable when applied to a specific subdomain (the field of medical surgery), and can be augmented to produce semantic information.

An advanced tagging-lemmatising tool will be developed in a bi-lingual environment as a stand-alone prototype. On acceptance of expressions or texts in natural language, the prototype will be capable of :

1. generating a lexicon where each word is tagged with syntactic information (category and subcategory) and a semantic labelling;
2. proposing alternative solutions in case of syntactic and/or semantic ambiguities;
3. ultimately providing translations of the tagged words in another European language;
4. facilitating the automatic classification and coding of medical information in a perspective of standardisation.

In order to lead the MULTI-TALE project to a definite success within a limited time scale of 18 months, with the involvement of three partners, a prototype tool will be built and tested on authentic sublanguage corpora, aiming at the automatic generation of coded mono- and multi-lingual specialised lexicons. This implies the following constraints :

1. the languages covered by the prototype will be Dutch and English. The overall approach however will guarantee portability to other European languages;
2. the domain of discourse will be limited to the sublanguage of surgical procedures, and more specifically the domain of neuro-surgery.

However, special attention will be devoted to develop a modular system in such a way that portability towards other domains of discourse in medical procedure (e.g. in the field of radiology, laboratory analysis, etc.) is guaranteed.

The MULTI-TALE proposal addresses **area II of the MLAP - Programme, i.e. the investigation of the suitability of advanced tools for lexicographic research and acquisition**.

MULTI-TALE contributes to the objectives and scope of the MLAP Programme in a number of ways:

- it is a response to the clear demand of having professionally designed and ready to use application-prototypes based upon widely available and commonly agreed methods, tools and resources;

- MULTI-TALE is in line with the ongoing activities of **standardisation** in the field of medical terminology and computational linguistics for medical sublanguage

- the proposal promotes the augmentation, the evaluation and use of NLP-tools in a key-application, i.e. the automatic (syntactic-semantic) tagging and generation of specialised lexicons in the field of medical surgery;

- the issue of multi-lingualism within the Community is addressed ;

- this proposal addresses fully the demand to bring together experts in various disciplines, each of them being able to build on existing knowledge and methods. Moreover, the Consortium consists of partners representing users, industry and academic researchers.

Aims and Scope

The primary aim of MULTI-TALE is to develop a syntactic-semantic tagger-lemmatizer that in a multi-lingual environment assists lexicologists in the field of health care informatics to classify automatically specialised terms in a standardisation perspective of the medical sublanguage. MULTI-TALE will therefore investigate how two existing grammatical lemmatizers-taggers (DILEMMA-2 and D-TALE), developed for general (i.e. domain-independent) language, are suitable when applied to a specific subdomain (the field of medical surgery), and can be augmented to produce (morpho)-semantic information.
An advanced tagging-lemmatising tool will be developed in a multi-lingual environment as a stand-alone prototype. On acceptance of expressions or texts in natural language, the prototype will be capable of :

1. generating a lexicon where each word is tagged with syntactic information (category and subcategory) and a semantic labelling according to the conceptual model elaborated by Project Team 002S of Technical Committee 251 of The European Centre for Standardisation (CEN\TC251\PT002S) [1, 2].

2. proposing alternative solutions in case of syntactic and/or semantic ambiguities;

3. ultimately providing translations of the tagged words in another European language;

4. facilitating the automatic classification and coding of medical information in a perspective of standardisation.

In this respect special attention will be devoted :

1. to develop a modular system in such a way that portability towards other domains of discourse in medical procedure (e.g. in the field of radiology, laboratory analysis, etc.) is guaranteed;

2. to use an existing tagger-lemmatizer developed for general (i.e. domain-independent) Dutch language, and augment it for tagging-lemmatising medical sublanguage, inclusive of a semantic tagging function ;

3. to use an existing tagger-lemmatizer developed and geared for tagging-lemmatising English medical sublanguage and augment it towards a semantic tagging tool.

4. to make use of existing conceptual frameworks for representing the semantics in medicine, as proposed by the CEN\TC251

5. to perform an empirical validation of the prototype by using it intensively in a real-world context.

With these issues in mind, the **main goal** of the MULTI-TALE project is:

to design a prototype of a multilingual tagger-lemmatizer based on DILEMMA-2 and D-TALE, that is able :

1. to take as input the sublanguage of neuro-surgical procedures expressed in English or Dutch;

2. to provide as output a lemmatised list of words with syntactic and semantic information, based on the semantic model of CEN/TC251/PT002S

3. to ultimately show the feasibility of translating this specialised lexicon from English to Dutch, and from Dutch to English, maintaining a unique and consistent semantic 'decoration' for each word-pair.

Secondary goals are:

1. to collect, structure and format authentic and representative corpora of surgical procedures expressed in natural language (English and Dutch) and to study relevant literature in the field of medical language tagging and semantic modelling;

2. to analyse the semantic model elaborated by CEN/TC251 applied to the field of surgical procedures;

3. to tag a statistically relevant lexicon of the sublanguage corpora with relevant syntactic-semantic information according to the recommendations of the CEN/TC251;

4. to pragmatically validate the semantic model underlying the subcategorization of the tagged lexicon and tune the tagging-lemmatising prototype in accordance with the principles and objectives of an existing classification and coding project at European level;

5. to integrate the sublanguage syntactic and semantic knowledge as formalised in two existing taggers-lemmatizers;

6. to integrate the tagging-lemmatising components for the Dutch and English medical sublanguage into a unique user-interface that will analyse the sublanguage input (i.e. text or expressions) and generate a corresponding multi-lingual specialised lexicon with the relevant syntactic-semantic information.

7. to use unanalysed parts of the corpora collected in 1. as test data to assess the performance of the MULTI-TALE tool, and to adapt it where needed.

8. to set up a pilot study to test the tagger-lemmatizers in a real-world environment.

9. to perform final modifications in line with the pilot study.

Project Coordinator:

Dr. W. Ceusters
Hazenakkerstraat 20
B-9520 Zonnegem (Sint Lievens Houtem)
Belgium
Tel: +32 53 62 24 57
Fax: +32 9 220 56 57

Partners

Prof. W. Martin, Free University Amsterdam, Lexicology Research Group
Dr. G. De Moor, RAMIT VZW, CEN\TC251 secretariat