

THE ZSTATION AS A COMPUTATIONAL AID FOR CORPUS BASED RESEARCH

Henri ZINGLÉ

LILLA
Université de Nice-Sophia Antipolis
BP-209 F-06204 Nice Cedex
henri@lilla.unice.fr

Abstract : The ZStation is a workbench for NLP based on linguistic ressources. Recently new functions were added to the system in relation with document retrieval which may be used for pure linguistic investigation or terminographic work on corpora : extraction of lexical units and in particular compound extraction (using linguistic specifications or a stochastic method), extraction of phraseologic units, collocation, generation of indices and concordances on the basis of words, lemmas morpho-syntactic categories, concepts, formants or conceptual graphs.

Keywords : corpus linguistics - document retrieval - terminography

The ZStation (Zinglé, 1994) was initially developped as a workbench for the development of resources for advanced natural language processing and related applications. It is based on the interaction of two levels : the INTELINGUISTIC LEVEL focusing on knowledge representation using conceptual graphs (Sowa, 1984) and the INTRALINGUISTIC LEVEL which takes into account the linguistic properties of natural languages (inflexion, derivation, composition, syntax etc...). The ZStation may be used to model linguistic knowledge in a broad range of natural languages.

The ZStation provides 3 major managers related respectively to resources, applications and tools. The RESOURCE MANAGER allows users to create and to test ontologies, dictionaries, inflexional and syntagmatic grammars as well as specification grammars tailored for applications. The APPLICATION MANAGER provides tools for lexical analysis and generation,

ISBN: 0 08 043 438X

automatic analysis of sentences into conceptual graphs, automatic generation of sentences from conceptual graphs, full text analysis, tools for document retrieval and terminologic work. The list of applications is not closed. The TOOL MANAGER offers a lot of possibilities for linguists which facilitate the development of linguistic resources (extraction of linguistic data from corpora, automatic building of dictionaries...). The ZStation is implemented in PDC-Visual Prolog 4.1 on the Windows platform.

The focus of the presentation will be put on the facilities offered by the ZStation for corpora investigation in relation with (1) terminographic work and (2) document retrieval.

1. THE ZSTATION AS A TOOL FOR TERMINOGRAPHIC WORK

The tool magnager of the ZStation provides three useful tools for terminographic work : extraction of simple lexical units, of compounds and of phraseological units. These tools use the same interface, which allows users to select the corpus files and the eventual required linguistic resources.

The extraction of simple units produces a file containing a sorted list of units with their occurrences. The result file may be edited or processed by any text editor or data processor running on the Windows platform.

For the extraction of compounds two different methods are provided. The first one is based on a stochastic method. The process looks for recurrent form sequences and build during text scanning a dynamic lookup structure, whose content is filtered and retrieved at the end of the process. Documents are processed in a single pass. The second method uses linguistic knowledge to extract patterns from the corpus. It is based on a specification grammar edited by users, which describes the formal structure of compounds. The algorithm is based on a top-down phrase structure grammar using chart parsing technique. Tokens are lemmatized during parsing. The output file provides a sorted list of word sequences associated to their occurrences in the corpus. Even if the second method is more attractive in a linguistic point of view, it is by far slower and less exhaustive than the previous one.

The extraction of phraseologic units is parallel to compound extraction, with the important difference, that tokens are not necessarily close together.

In order to spare time in dictionary building a tool was developed to build dictionaries directly from corpora. The method uses a referent dictionary : recognized words are inserted after lemmatization in the new dictionary with all the related intralinguistic information (ontological, morpho-syntacical and structural links); unknown words are inserted in the new dictionary with a sign for later verification. New dictionaries may also be incremented automatically by the lookup of new documents. In practice machine assisted dictionary building reveals to be a very fast way to to build domain dictionaries.

2. THE ZSTATION AS A TOOL FOR DOCUMENT RETRIEVAL

The ZStation may also be used to build indexes, in particular : ◊ *word indexes* : this process registers all document references for words (it needs a dictionary for lemmatization) ; ◊ *category indexes* : this process allows to search all adjectives in the specified document for example ; ◊ *morphem indexes* : this process looks for all words containing morphems

ISBN: 0 08 043 438X

specified by users (prefixes, suffixes, infixes, stems) ; ◊ *concept indexes* : this process looks for all words belonging to a given concept or a conceptual relation (it combines the use use of dictionaries and of an ontology with its conceptual inference engine).

A dialog allows users to select the corpus and the required linguistic resources. Indexes may be accessed using a query interface or appropriate data converters.

The query interface provides the list of index items. The information related to a selected item will appear in a text field of the query interface with the indication of the reference (source document and position within it). Users may select samples and paste them in another text field, which may be printed directly or saved to a file. This interface is very useful for corpora investigation in general and also for the validation of terms and phraseologic units in terminology.

Indexes are databases with encoded data. Their content is not directly readable and has to be converted for non interactive use. Users may choose to convert them either into statistics or into a concordance. The first function lists all items in alphabetical order with the indication of the sum of references within individual files as well as in the whole corpus. The header of the file contains the directory of all processed files of the corpus. The second function creates a similar file, where each item is associated to all sentences in which it occurs. To each sentence are associated the reference of the document it belongs to and the position within the document. In both cases the result file may be edited or processed by commercial applications of the Windows platform.

The ZStation offers in addition an application for interactive full text analysis. A document retrieval language was developed using powerful commands which users may combine to scan texts. To look for example for a noun indicating an animal which is subject of a verb indicating an aerial travel mode the following combination of commands should be used : {match(N,V),cat(N,sub),cat(V,vb),cpt(N,+isa,animal0),cpt(V,+isa,fly0),accord(sv,N,V)}.This method is very useful to retrieve information by combining morphology, syntax and semantics. It is less easy for users as it requires to be familiar with the formalism used in the ZStation for syntagmatic grammars and ontologies.

An application for full text analysis using conceptual graphs is under development. The goal here is to locate information, even if it is expressed in different ways. Good results were obtained for the extraction of information in nominal vs verbal structures and in passiv vs activ structures. The method applies now to any kind of structures. The goal is to translate the query into conceptual graphs (in a precise knowledge domain) and to map the information found in documents with the query representation.

Remarks

The ZStation is actually used in partner research teams either for corpora investigation (like in the GTM project at the university of Granada in Spain) or for terminography (like in the LEXTERM project at the university of Brasilia). It is intensively used in our research team in the project TASTE which is oriented to the linguistic analysis of documentation in science and technology.

ISBN: 0 08 043 438X

REFERENCES

Sowa, J. (1984) *Conceptual structures : information processing in man and machine.* Addison-Wesley.

Zinglé, H. (1994) The ZStation workbench and the modelling of linguistic knowledge. *Current issues in mathematical linguistics*, Elsevier-NHLS, 1994, 423-432

ISBN: 0 08 043 438X