Lemmatizer

Fri, 16/05/2008 - 16:25 — webmaster

Description

Lemmatizer is the language tool that given any word-form it returns the headword of the morphological lemma this word-form belongs to, e.g. for the word-form *κατέστη* it returns the headword *καθιστώ*. In cases where the input word-form belongs to more than one morphological lemmata, the Lemmatizer returns all the corresponding headwords, e.g. for the word-form *απαντήσεις* it returns *απαντώ* and *απάντηση*. The functionality of Neurolingo's Lemmatizer is based on an index that contains all the word-forms of the Morphological Lexicon [1].

For an extensively inflected language like Modern Greek, the Lemmatizer is an integral functional component of text indexing and searching systems. For example, when the user searches for texts containing the word-form *υπολογιστές*, he/she would probably like to recall texts that also contain the word-forms *υπολογιστής*, *υπολογιστή* and *υπολογιστών*. In practice, this is achieved by lemmatizing the indexing terms as well as the search terms: texts that contain the word-forms *υπολογιστή*, *υπολογιστές* and *υπολογιστών* will also be indexed with the headword *υπολογιστής*; a query that contains the word-form *υπολογιστές* will be expanded with the headword *υπολογιστής*. This way, a text containing the word-form *υπολογιστών* can now be associated with a query containing the word-form *υπολογιστές* through the common headword *υπολογιστής*.

Try Lemmatizer online. [2]

Applications

The Lemmatizer of Neurolingo has been intergrated with the following text indexing and searching systems:

- Apache Lucene [3]. The functionality of the Lemmatizer is offered through a descendant of the Java class org.apache.lucene.analysis.Analyzer.
- Oracle Text [4]. The functionality of the Lemmatizer is offered through stored procedures (USER_LEXER preference).
- Microsoft Indexing Service [5] / SQL Server Full-Text Search [6]. The functionality of the Lemmatizer is offered through the implementation of the IStemmer COM interface.

Source URL: http://www.neurolingo.gr/en/technology/application_tools/lemmatizer.jsp

Links:
[1] http://www.neurolingo.gr/en/technology/lexica/morpholexicon.jsp
[2] http://www.neurolingo.gr/en/online_tools/lexiscope.htm
[3] http://lucene.apache.org
[4] http://www.oracle.com/technology/products/text/index.html
[5] http://msdn2.microsoft.com/en-us/library/bb187804.aspx
[6] http://msdn2.microsoft.com/en-us/library/ms345119.aspx