DIAC+

1. BASIC INFORMATION
    1. Tool name

    DIAC+: A Professional Diacritics Recovering System

    2. Overview and purpose of the tool

    We propose an accurate knowledge-based system for automatic recovering the missing diacritics in MSOffice documents written in Romanian. The out of lexicon words (usually, very rare) are processed by a character-based back-off procedure.For the rare cases when the system is not able to reliably make a decision, it either provides the user a list of words with their recovery suggestions, or probabilistically choose one of the possible changes, but leaves a trace (a highlighted comment) on each word the modification of which was uncertain.

    3. A short description of the algorithm

    Since the DIAC+ was designed to work with MS formatted documents, the system extracts the textual data from the input file and stores it in an internal format adequate for our pre-processing tools, using as database a full-text search engine – Lucene3. The textual data extracted from the input file is tokenized and tieredtagged, thus creating a linguistic knowledge space for the current text within which the proper restoration of diacritics takes place.

    In the hypotheses generation step, a word is first searched in the union of D0 and D1 dictionary because in a text without diacritics or with partial diacritics one cannot be sure if a word is in its regular form or not unless contextual information is available.

    If the word cannot be found in the union of D0 with D1 it is searched in the D2 dictionary. A word which is not found in any of the system's lexicons is considered unknown and irrecoverable by the word-based approach, and its processing is left in charge of a character-based recovery module.

2. TECHNICAL INFORMATION
    1. Software dependencies and system requirements

    Microsoft Word 2007 or 2010; the configuration programs automatically downloads and installs Windows Installer, .Net Framework 3.5 and VS Tools for Office; internet connection for installation and update;

    2. Installation

    We do not have a third-party SSL certificate and the program is installed and updated using a RACAI-generated SSL certificate. To accept it, you need to add http://www.racai.ro as a trusted site:

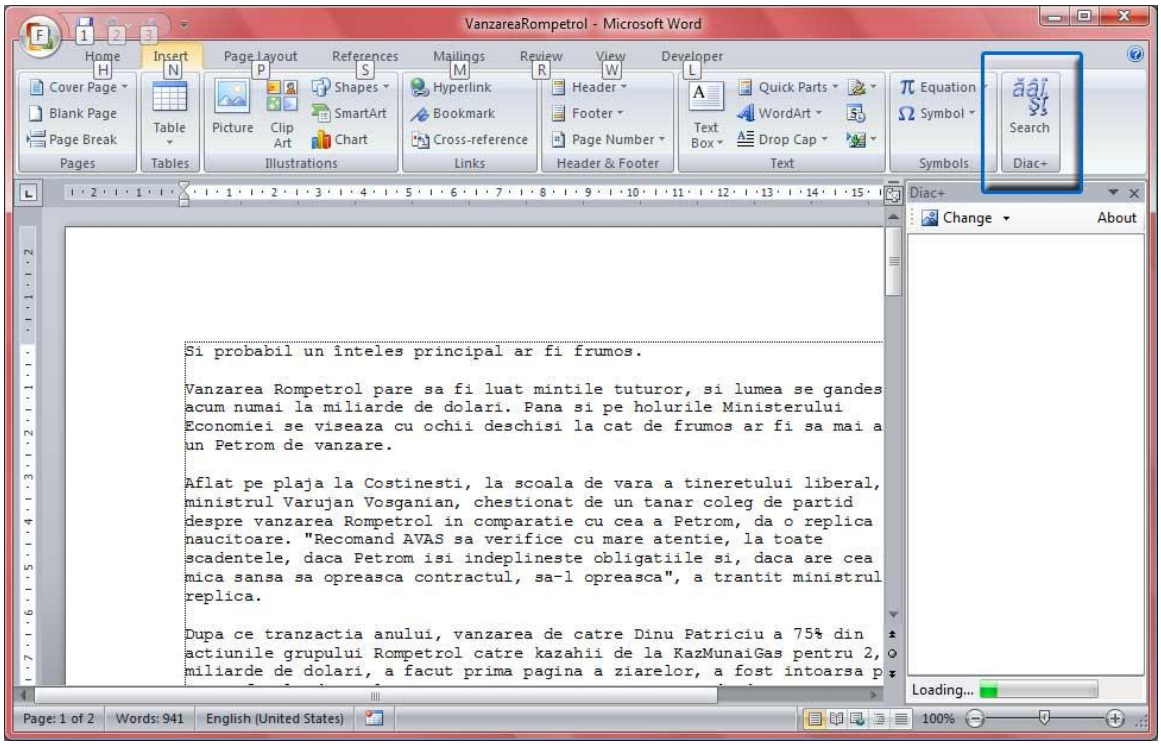
    Invoke *Tools* under *Internet Explorer*, choose the menu option *"Internet Options"* from the *"Internet Options"* window and select the *"Security"* tab. Then select *"Trusted Sites"* and push the button *"Sites"*. A window called *"Trusted sites"* will pop-up and you can add the site

http://www.racai.ro and uncheck the *"Require server verification..."* option.

Now you can install the application by following this link: http://www.racai.ro/diac/downloads/setup.exe and executing the setup program or by executing the setup.exe from the Installation folder. DIAC+ will be installed as a plugin for Microsoft Word 2007 or 2010.

3. Execution instructions

Load a MS office document and press the DIAC+ button, installed automatically in the Insert toolbar.



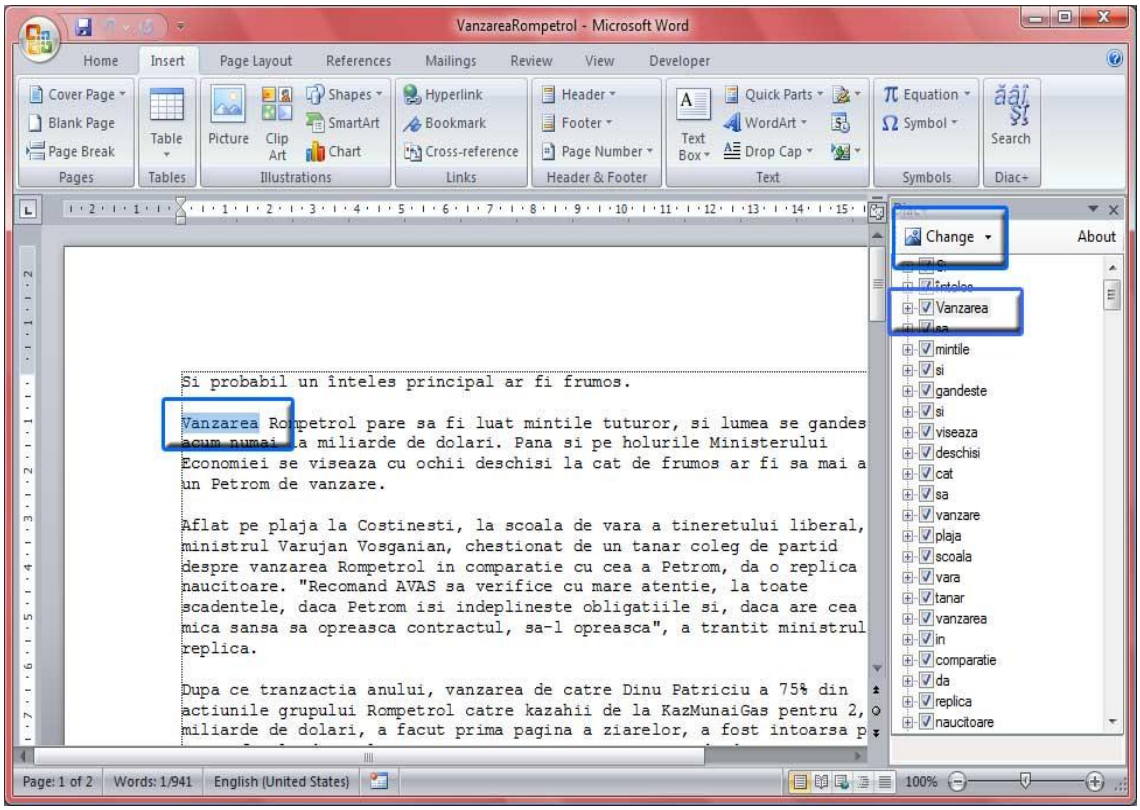
4. Input/Output data formats

a. Input data formats

MS Word 2007 or 2010 documents.

b. Output data formats

The program offers a list of suggestions, some of them already checked for the cases were the application considers there is little or no ambiguity. You can accept the suggestions by double-clicking them or by pushing the button *Change.* There is a Change All button also if you open the Change Menu. For the ambiguous cases, the user must choose between the variants proposed by DIAC+.



5. Integration with external tools

Easy and user-friendly integration as a plugin in MS Word.

3. CONTENT INFORMATION
    1. a test input file

    See the text_input.docx file in the '`test/`' directory.
    2. the output file

    See the output docx file in the '`test/`' directory.
    3. approximation of the time necessary to process the test input file.

    Intel I7, 3.3Ghz: 8 seconds.

4. ADMINISTRATIVE INFORMATION
    1. Contact

    Tufiș Dan tufis@racai.ro "Mihai Drăgănescu" Research Institute for Artificial Intelligence of the Romanian Academy – department: NLP Group

5. RELEVANT REFERENCES AND OTHER INFORMATION

Tufiş, D., Ceauşu. D. (2008)*DIAC+: A Professional Diacritics Recovering System*. In*Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0

Tufiş, D., Ceauşu, A. (2007):*Diacritics Restoration in Romanian Texts*. In Elena Paskaleva and Milena Slavcheva (eds.), *A Common Natural Language Processing Paradigm for Balkan Languages - RANLP 2007 Workshop Proceedings*, pp. 49-56,

Borovets, Bulgaria, September 2007. INCOMA Ltd., Shoumen, Bulgaria. ISBN 978-954-91743-8-0

Tufiş, D., Ceauşu, A. (2007):*DIAC+: Un sistem profesional de recuperare a diacriticelor*. In Ionuţ Pistol, Dan Cristea, and Tufiş, D. (eds.), *Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române*, pp. 151-160, Iaşi, Romania, dec. 2007. Universitatea "Al.I. Cuza" Iaşi, Editura Universităţii "Al.I. Cuza" Iaşi. ISBN 978-97-3703-297-3

Tufiş, D., Chiţu, A. (1999):Automatic Insertion of Diacritics in Romanian Texts. In *Proceedings of the 5th International Workshop on Computational Lexicography COMPLEX*, Pecs, Ungaria, 1999, pp. 185-194