CST's lemmmatiser - Greek version

Bart Jongejan
CST-University of Copenhagen
Njalsgade 140-142
DK-2300 Copenhagen S
bartj@hum.ku.dk

1 Description

CST's lemmatiser (Jongejan and Dalianis, 2009) analyses not just the endings of words for suffixes that undergo change under lemmatisation, but also prefixes and infixes, if necessary.

The material for training of the greek lemmatisation rules was kindly provided by George Petasis. (http://www.ellogon.org/petasis/). See Petasis et al. (2001)[1], Petasis et al. (2003)[2]

Use the program 'cstlemma' to lemmatize greek word forms using the file 'flexrules' as linguistic resource by providing the command line option '-f flexrules'. This flex rule file lemmatizes al words in the training set correctly. We estimate that 10 percent of OOV words are lemmatised incorrectly, but correctness may become much more unfavourable if the input text has no diacritics.

The most recent version of the lemmatizer can be downloaded from https://github.com/kuhumcst/cstlemma.

References

Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153. Association for Computational Linguistics.

Georgios Petasis, Vangelis Karkaletsis, Dimitra Farmakiotou, Ion Androutsopoulos, and Constantine D. Spyropoulos. 2001. A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker. In *Proceedings of the 8th Panhellenic Conference on Informatics (PCI'01)*, PCI'01, pages 80–89, November 8–10.

Georgios Petasis, Vangelis Karkaletsis, Dimitra Farmakiotou, Ion Androutsopoulos, and Constantine D. Spyropoulos. 2003. A Greek Morphological Lexicon and Its Exploitation by Natural Language Processing Applications. In Yannis Manolopoulos, Skevos Evripidou, and Antonis Kakas, editors, *Advances in Informatics - Post-proceedings of the 8th Panhellenic Conference in Informatics*, volume 2563 of *Lecture Notes in Computer Science*, pages 401–419. Springer Berlin / Heidelberg. http://www.springerlink.com/content/hcdjrlvj5nlybf5c/.

[1] http://www.ellogon.org/petasis/bibliography/PCI2001/EPY-Morph-CameraReady.pdf

[2] http://www.ellogon.org/petasis/bibliography/PCI2003/25630398.pdf