MLRS THESAURUS

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

This is an automatically produced distributional thesaurus, which finds words that tend to occur in similar contexts as the target word. It is not a manually constructed thesaurus of synonyms. It was produced on the basis of the MLRS corpus.

1.2 Representation of the lexicon (flat files, database, markup)

The thesaurus consists of a single text file: thes_mt02.txt

1.3 Character encoding

UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Adam Kilgarriff
http://www.sketchengine.co.uk/
adam@lexmasterclass.com

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

Text file for download within META-SHARE.

2.3 Copyright statement and information on IPR

3. TECHNICAL INFORMATION

3.1 Directories and files

Single text file: thes_mt02.txt

3.2 Data structure of an entry

Each entry is on one line, containing the lemma frequency and the lemma in first two columns, and the rest of the line is a list of up to 20 best matches pairs (lemma+score) that have a score of at least 0.2.

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

36,034 entries; 9.7 MB (9.649.345 Bytes) occupied on disk

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Maltese

4. 2 Entry Type

One line per lemma, with tabs separating lemma frequency, lemma and up to 20 best matching pairs (lemma+score).

For example, the entry for *xogħol* „work“ looks like this in the text file:

```
178918 xogħol      servizz_0.290413796902  proġett_0.271775722504
ħaddiema_0.263248413801 ħidma_0.247022345662
xogħlijiet_0.246716752648     gvern_0.243715748191
post_0.238192170858     affarijiet_0.236238643527
servizzi_0.236086919904 programm_0.232582449913
nies_0.226817399263     proċess_0.221314221621  mod_0.216820463538
saħħa_0.216409176588    malta_0.215376183391
sena_0.213514536619     ħin_0.212257444859
pajjiż_0.210465654731   ħajja_0.210368424654    każ_0.210191175342
```

4.3 Attributes and their values

lemma_freq (numeral)
lemma (string)
match pair (lemma(string)_score(numeral))

4.4 Coverage of the lexicon

All the text domains within the MLRS corpus.

4.5 Intended application of the lexicon

Dictionary building and research

4.6 POS assignment

N.A.

4.7 Reliability (automatically/manually constructed)

This is an automatically produced distributional thesaurus, which finds words that tend to occur in similar contexts as the target word. It is not a manually constructed thesaurus of synonyms. It was produced on the basis of the MLRS corpus.

5. RELEVANT REFERENCES AND OTHER INFORMATION

http://www.sketchengine.co.uk