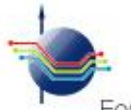

Ερευνητικό Κέντρο Αθηνά
Athena Research Center

# Automatic recognition and extraction of multiword nominal expressions from corpora

**Angeliki Fotopoulou, Giorgos Giannopoulos , Maria Zourari, and Marianna Mini**

**Institute for Language and Speech Processing, "Athena" RIC**

afotop@ilsp.athena-innovation.gr giann@imis.athena-innovation.gr

**Cost PARSEME**

**WP2**

# Method

Ερευνητικό Κέντρο Αθηνά
Athena Research Center

Nominal multiwords: *παιδική χαρά* [lit. "kids' joy", meaning "playground"], *ψήφος εμπιστοσύνης* ["vote of confidence"]

Tools Corpus : ~142.000.000 words (930 MB)

ILSP morphological electronic dictionary

Ngram Statistics Package tool (Pedersen et al., 2006)

Method

(a) Six (6) general grammar rules are applied to a grammatically tagged corpus

(b) Results are filtered using more specific grammar rules and filters (word lists)

(c) Results are evaluated using statistical methods

(d) A linguist/encoder makes the final selection and MWE are stored in a Database

# Results

| | Adj + N | N + N(gen.) | N + N(gen.) | N + - + N |
|---|---|---|---|---|
| 1st Step: Application of 6 general rules | 1.068.000 | 242.000 | 619.000 | 120.000 |
| 2nd Step: Application of specific rules filters | 570.000 | 189.000 | 347.000 | 12.000 |
| 3rd Step: Application of statistical metrics | 68.000 | 22.000 | 31.000 | 400 |

# Database for MWEs

Ερευνητικό Κέντρο Αθηνά
Athena Research Center