



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

File Extraction

Επεξεργασία Πολυλεκτικής Έκφρασης

Έκφραση:

Δομή:

Κλίση: 1ο Συστατικό 2ο Συστατικό

Ενικός: Ονομαστική
 Γενική
 Απασκή
 Κλητική
Πληθυντικός: Ονομαστική
 Γενική
 Απασκή
 Κλητική

φαι νόμνο	του	θερμοκηπίου	ου	θερ
φαι νομένου	του	θερμοκηπίου	ου	θερ
φαι νόμνο	του	θερμοκηπίου	ου	θερ
-	-	-	-	-
φαι νόμνα	των	θερμοκηπίου	ων	θερ
φαι νομένων	των	θερμοκηπίου	ων	θερ
φαι νόμνα	των	θερμοκηπίου	ων	θερ

Ονοματοποιείται:

Κατηγορία:

Εναλλακτικό:

Είναι πραγματικά:

Επίπεδο γλώσσας:

Είδος:

Σχετικά Ρήματα

Νέο ρήμα

προκαλώ	66
είμαι	25
συμβάλλω	21
οφείλω	9

Στοιχεία κλίσης

φαι νόμνο του θερμοκηπίου	373
φαι νομένου του θερμοκηπίου	50
φαι νόμνου του θερμοκηπίου	2
φαι νόμνα του θερμοκηπίου	1

Σχόλια

Λίστα Πολυλεκτικών Εκφράσεων

Προβολή διεγραμμένων εκφράσεων -Σταp
 Προβολή αποθηκευμένων εκφράσεων
 Εμφάνιση λίστας υπαρχόντων πολυλεκτικών

Κατηγορία:

Επίπεδο γλώσσας:

Είδος:

Αναζήτηση: Ακρβής λέξη

1. φας ο δημοσιότητα
2. αξία ο συναλλαγή
3. διάταξη ο άρθρο
4. συνήγορος ο πολίτης
5. αναθεώρηση ο σύνταγμα
6. προστασία ο περιβάλλον
7. κάδωνας ο κίνδυνος
8. λωρίδα ο γάζα
9. μερίς ο λέων
10. κατάσταση ο υγεία
11. αιχμή ο δόρυ
12. υποτίμηση ο δραχμή
13. κέντρο ο πόλη
14. όγκος ο συναλλαγή
15. διανομή ο κέρδος
16. φαινόμενο ο θερμοκηπιο
17. μήλο ο έριδα
18. επιστροφή ο μάρμαρο
19. πέραςμα ο χρόνος
20. πόλεμος ο κόλλο
21. πάροδος ο χρόνος
22. συμβολή ο οδός
23. συγγενής ο θήμη