

"Ιδιώματα, αυτά τα ατίθασα..."

Μια συγκριτική ανάλυση μεταξύ
λογοτεχνικών σωμάτων
κειμένων και Twitter με τη βοήθεια
του εργαλείου MWETOOLKIT3 για
την ανακάλυψη ρηματικών
πολυλεκτικών εκφράσεων.



Μάλλη Μαριλένα
Μαρκαντωνάτου Στέλλα
Ξυλογιάννη Άρτεμις
Στάμου Βίβιαν
Τακορού Πέννυ

ΑΘΗΝΑ

2020

[illegible]

ΣΤΟΧΟΣ ΤΗΣ ΔΟΥΛΕΙΑΣ ΜΑΣ

- Δοκιμή του MWEtoolkit3 για πρώτη φορά για συλλογή ελληνικών ρηματικών ΠΛΕ

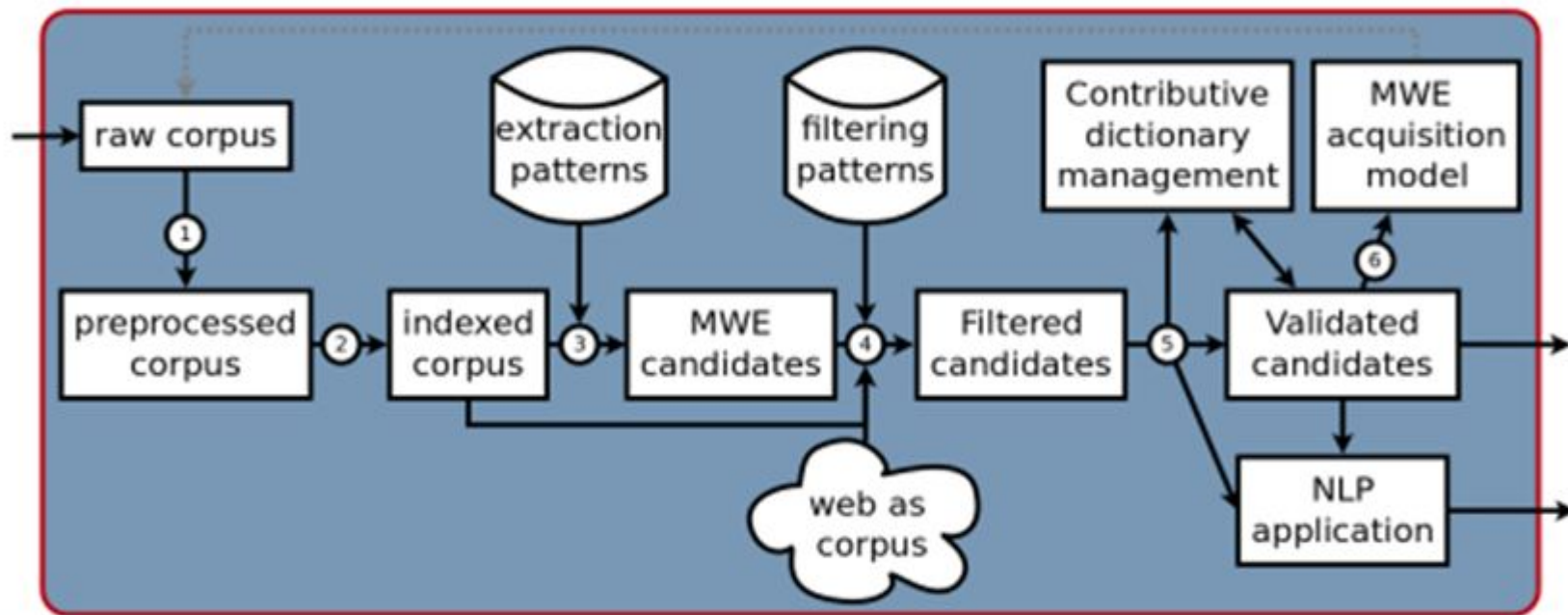
- Εμπλουτισμός της λεξικογραφικής βάσης δεδομένων του IDION με ρηματικές ΠΛΕ της Νέας Ελληνικής Γλώσσας

(Markantonatou et al. 2019)

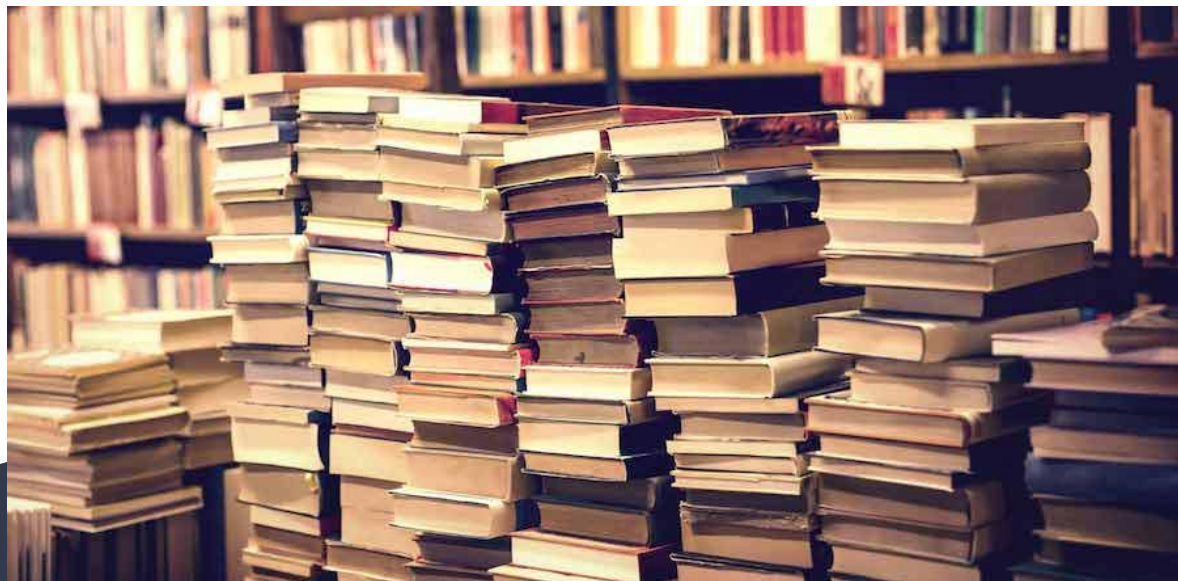
<http://idion.ilsp.gr/>



ΠΩΣ ΔΟΥΛΕΥΕΙ ΤΟ ΕΡΓΑΛΕΙΟ ΜΑΣ



Fiction corpora



CORPUS

➤ Αποτελείται από 5 υποσώματα:

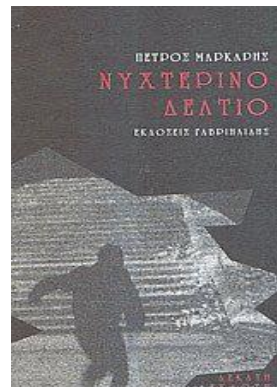
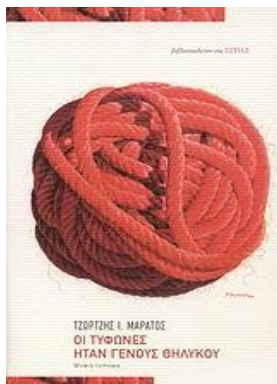
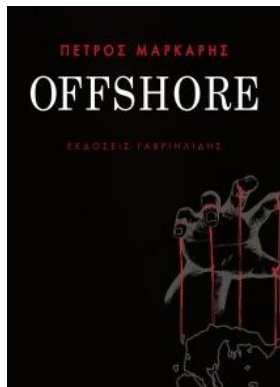
Μάρκαρης (2016) *Offshore*

Μαράτος (2007) *Οι τυφώνες ήταν γένους θηλυκού*

Παπαδάκη (2001) *Η βαρκάρισσα της Χίμαιρας*

Μάρκαρης (1995) *Νυχτερινό δελτίο*

Ταχτσής (1970) *Τρίτο στεφάνι*



ΚΟΙΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

- Εκδόθηκαν μετά το 1970
- Είναι πλούσια σε ρηματικές πολυλεκτικές εκφράσεις
- Καθημερινός λόγος - Προφορικότητα

1ο ΣΤΑΔΙΟ: ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΑΓΡΑΦΗΣ

- Οι επισημειωτές διαβάζουν τα βιβλία
- Καταγράφουν πιθανές VMWEs
- Η διαδικασία επαναλαμβάνεται 2 φορές από 2 διαφορετικούς annotators
- Interannotator Agreement
- Προέκυψε ωστόσο η ανάγκη 3ου annotator (λόγω χαμηλής συμφωνίας μεταξύ των 2 πρώτων)





Comparison_Maratos



Αρχείο Επεξεργασία Προβολή Εισαγωγή Μορφή Δεδομένα Εργαλεία Πρό

Σε εξέλιξη...

Η τελευταία τροποποίηση



100%



.0

.00

123

Arial



10



fx

	A	B	C	D
1	(Markantonatou)	(Penny) - Annotator A	(Marilena) - Annotator B	(Markantonatou)
2		1 «...δίνουν άλλο χρώμα στο χωριό..»	«...σαν να της έδινε διαταγή...»	1
3		1 «... ήταν κυριολεκτικά να την πεις στο ποτήρι»	«...περνούν την ώρα τους...»	1
4		1 «...την έτρωγε από την κορφή μέχρι τα νύχια.»	«...να της δώσουν έντολές...»	1
5		1 «... την έγδυne με τα μάτια του...»	«...να γίνει το δικό μας...»	1
6		1 «... δεν ξεκολλούσε τα μάτια του από πάνω της.»	«...δεν έπαιρνε και τόσο κατακαρδα...»	1
7		1 «... χτυπημένος στο δόξα πατρί»	«...και μαύριζε η καρδιά του...»	1
8		1 «έβαζαν με το μυαλό τους κάτι ιδιαίτερο»	«...να βγάλει λεφτά, να γίνει ανεξαρτητος...»	1
9		1 «το ειδύλλιον πλέχτηκε»	«...θα τα έβρισκε μπροστά του και θα τα ζούσε ο ίδιος...»	1
10		1 «θα πάρω το πρώτο αεροπλάνο»		
11		«μου το χρωστάει»	«...στηριζόταν πάνω του...»	
12		1 «Εγώ θα τον κάνω άντρα »	«...άνοιξε την καρδιά της...»	1
13		1 «...έκαναν καλή παρέα. »	«...έπρεπε να του δώσει τη χυλόπιτα...»	1
14		1 «Είσαι με τα καλά σου ; »	«...να βγάλω τα εισιτήρια...»	1
15		1 «είχε αφήσει το βλέμμα του να πλανιέται»	«...έβγαλα τρία εισιτήρια...»	1
16		«...είχαν θολώσει το μυαλό του για τα	«...δεν μου άρεσε στο πρώτο μου ταξίδι	



Annotators

common MWEs

non common MWEs

All MWEs

MWEs gia excel



100% € % .0 .00 123 Προεπιλογ... 10 B I A

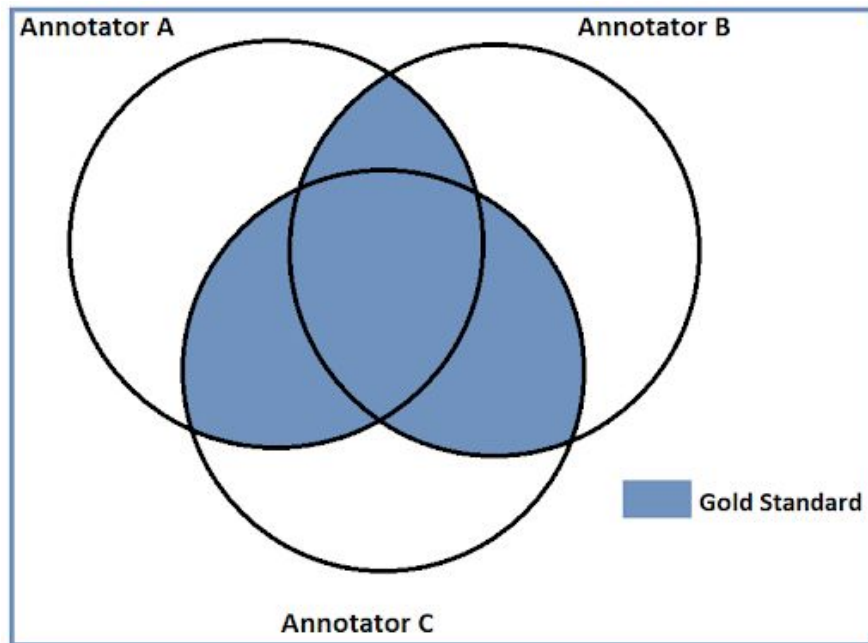
	A	B	C	D	E
1	Offshore-Μάρκαρης	Νυχτερινό Δελτίο-Μάρκαρης	Μαράτος	Παπαδάκη	Ταχτσής
2	Nc = 2 raters Ni = 748 items Nk = 714 categories Nc x Ni = 1496 judgements Overall agreement coefficients for all annotators: multi-ao = 0.878342 multi-pi (Fleiss' kappa) = 0.877989 multi-kappa = 0.878150 Weighted agreement coefficients for all annotators: alpha = 0.878071 alpha-kappa = 0.878150	Nc = 2 raters Ni = 1196 items Nk = 1148 categories Nc x Ni = 2392 judgements Overall agreement coefficients for all annotators: multi-ao = 0.933946 multi-pi (Fleiss' kappa) = 0.933818 multi-kappa = 0.933889 Weighted agreement coefficients for all annotators: alpha = 0.933845 alpha-kappa = 0.933889	Nc = 2 raters Ni = 591 items Nk = 556 categories Nc x Ni = 1182 judgements Overall agreement coefficients for all annotators: multi-ao = 0.939086 multi-pi (Fleiss' kappa) = 0.938914 multi-kappa = 0.938972 Weighted agreement coefficients for all annotators: alpha = 0.938966 alpha-kappa = 0.938972	Nc = 2 raters Ni = 621 items Nk = 659 categories Nc x Ni = 1242 judgements Overall agreement coefficients for all annotators: multi-ao = 0.772947 multi-pi (Fleiss' kappa) = 0.771232 multi-kappa = 0.772657 Weighted agreement coefficients for all annotators: alpha = 0.771417 alpha-kappa = 0.772657	Nc = 2 raters Ni = 1046 items Nk = 976 categories Nc x Ni = 2092 judgements Overall agreement coefficients for all annotators: multi-ao = 0.974187 multi-pi (Fleiss' kappa) = 0.974127 multi-kappa = 0.974131 Weighted agreement coefficients for all annotators: alpha = 0.974140 alpha-kappa = 0.974131
3					
4					
5	Gold Standard: common elements between annotatorA(a+b) and annotatorB				
6					
7	Initials Meaning				
8					
9					
10					

ΚΡΙΤΗΡΙΑ ΚΑΤΑΓΡΑΦΗΣ

- Επισημειωτές : Φυσικοί ομιλητές της γλώσσας
- Διαισθητικά
- Με βάση τη συχνότητα που το συναντήσαμε στο σώμα κειμένων
- Με βάση τη συχνότητα που το συναντήσαμε στο διαδίκτυο

GOLD STANDARD

- Συνολικά προέκυψαν 3,500 ρηματικές πολυλεκτικές εκφράσεις



corpus.txt

- 1 τόσο του κόβει
- 2 τρέχουμε και δε φτάνουμε
- 3 Ρίχνω ένα βλέμμα
- 4 ρίξει μια ματιά
- 5 μπήκε στον κόπο
- 6 κάνουν το κομμάτι τους
- 7 Έσκασε μύτη
- 8 ήταν χαμένη απο χέρι
- 9 σου σπάει τα νεύρα
- 10 κλείσω ραντεβού
- 11 με πιάνει το πείσμα μου
- 12 παίρνεις σβάρνα
- 13 μου κάθονται στο στομάχι
- 14 κάνω τράκα
- 15 Ρίχνει ένα φευγαλέο βλέμμα
- 16 μπαίνει αμέσως στο νόημα
- 17 έπεσα μέσα
- 18 τον σπάσουν στο ξύλο
- 19 ιδρώσει ούτε τ' αφτί του
- 20 Ρίχνει ένα βλέμμα έξω
- 21 ρίξει κανένας τρίτος μια ματιά
- 22 έπαιρνε στο τηλέφωνο
- 23 πήγε στράφι

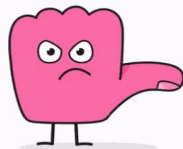
ΤΙ ΠΕΡΙΛΑΜΒΑΝΕΙ ΤΟ CORPUS ΜΑΣ :

- Όλο το corpus μας αποτελείται από **ρηματικές ΠΛΕ** («κιτρίνιζε από το κακό του» / «Αντί να βάλει την ουρά κάτω απ' τα σκέλια» / «Έσκασε μύτη...» / «...κάνω τράκα...»/ «Δεν ήταν παίξε-γέλασε»)
- **Proverbs** («Ήταν ξερό το κλήμα το 'φαγε κι ο γάιδαρος»)
- **Stereotyped similes** («...ξεγλιστράνε σαν το χέλι»/ «έτρεμε σαν το ψάρι» / «κοκκίνιζε σαν το παντζάρι»)
- **Routine formulas** («Πώς πάει;»)

2ο ΣΤΑΔΙΟ: ΑΠΑΡΑΙΤΗΤΗ ΔΙΑΔΙΚΑΣΙΑ ΠΡΙΝ ΤΗ ΧΡΗΣΗ ΤΟΥ Mwetoolkit3

- Scanning
- Optical Character Recognition (OCR)
- Διόρθωση λαθών OCR





ΠΕΤΡΟΣ ΜΑΡΚΑΡΗΣ

«Να μην ψάξουμε για λεφτά;» με ρώτησε ο Σωτήρης, ο υπαστυνόμος, που είναι ψείρας.

«Ψάξε κι ό,τι βρεις δικιά σου, αλλά δε θα βρεις δεκάρα. Είτε γιατί δεν είχαν είτε γιατί τα πήρε ο φονιάς. Χωρίς αυτό να σημαίνει ότι τους σκότωσε για να τους ληστέψει. Και για εκδίκηση να τους σκότωσε, τα λεφτά θα τα 'παίρνε έτσι κι αλλιώς. Δεν υπάρχει περίπτωση να βρουν αυτοί λεφτά και να τ' αφήσουν». Έψαξε και βρήκε μια τρύπα στο στρώμα. Λεφτά δε βρήκε.

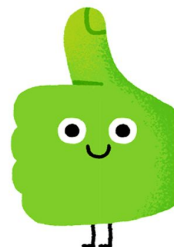
Από τους γείτονες, κανείς δεν είχε δει τίποτα. Έτσι έλεγαν τουλάχιστον. Τώρα, μπορεί και να το έκρυβαν για να το πουν μπροστά στην κάμερα και να κάνουν το κομμάτι τους. Δεν έμενε παρά να πάμε στην υπηρεσία για τη δεύτερη φάση – μια αναφορά που θα πήγαινε κατευθείαν στο αρχείο. Γιατί να ψάχνουμε αυτόν που τους σκότωσε θα ήταν χαμένος κόπος.

Έσκασε μύτη πάνω που σφραγίζαμε το σπίτι. Φεγγαροπρόσωπη, με μια μπλούζα φωσφορίζουσα, που τα στήθια της πάλευαν να την ξεσκίσουν για να πεταχτούν έξω μια φούστα στενή και πιο κοντή πίσω, γιατί ο κώλος τη την εμπόδιζε να κατέβει, και λιλά πασούμια. Καθόμου στο περιπολικό, όταν την είδα να πηγαίνει στους δύο άντρες που σφραγίζουν την πόρτα. Κάτι τους ψιθύρισε και αυτοί έδειξαν εμένα. Πήρε στροφή και με πλησίασε.

«Πού μπορώ να σας μιλήσω;» με ρώτησε, λες και περίμενε να της κλείσω ραντεβού.

«Εδώ. Λέγε».

«Τις τελευταίες μέρες είδα έναν άντρα να γυροφέρνει το σπίτι. Χτυπούσε να του ανοίξουν, αλλά η γυναίκα του έκλεινε κάθε φορά την πόρτα στα μούτρα. Ήταν ένα μέτριος, ξανθουλός, με μια ουλή στο αριστερό μάγουλο. Φορούσε ένα γαλάζιο μπουφάν, ένα τζιν μανταριαμένο στο γόνατο και αθλητικά παπούτσια. Τελευταία φορά το



180

Κάθε πρωί στις εννιά κοιταζόμαστε. Αυτός είναι όρθιος μπροστά στο γραφείο μου, με το βλέμμα κολλημένο πάνω μου, όχι ακριβώς στο ύψος των ματιών, αλλά λίγο πιο ψηλά, κάπου εκεί ανάμεσα στη βάση του μετώπου και στις βλεφαρίδες. «Είμαι μαλάκας» μου λέει.

Όχι με λόγια, με το βλέμμα του μου το λέει. Εγώ είμαι καθισμένος πίσω από το γραφείο και τον κοιτάζω ακριβώς στην κόρη, ούτε πιο ψηλά ούτε πιο χαμηλά. Γιατί εγώ είμαι ο προϊστάμενός του και μπορώ να τον κοιτάζω στα μάτια, αυτός δεν μπορεί. «Το ξέρω ότι είσαι μαλάκας» του λέω. Ούτε εμένα βγαίνει λέξη από το στόμα μου, το βλέμμα μου μιλάει. Αυτή τη συζήτηση την κάνουμε δέκα μήνες το χρόνο, αν εξαιρέσεις τους δύο που είμαστε ο καθένας σε άδεια, πέντε μέρες την εβδομάδα, Δευτέρα ως Παρασκευή, χωρίς να λέμε λέξη, μόνο με το βλέμμα. «Είμαι μαλάκας» – «Το ξέρω ότι είσαι μαλάκας».

Σε κάθε υπηρεσία αναλογεί κι ένα ποσοστό αποτυχιμένων. Δεν μπορείς να 'χεις μόνο σαΐνια, θα φορτωθείς και μερικούς στούρνους. Ο Θανάσης ανήκει σ' αυτό το ποσοστό. Μπήκε στη Σχολή Αξιωματικών της Αστυνομίας, αλλά τα παράτησε στη μέση. Με τα χίλια βάσανα κατάφερε να φτάσει αρχιψύλακας, κι εκεί άραξε. Καμία φιλοδοξία για παραπάνω. Από την πρώτη μέρα που ήρθε στην υπηρεσία, φρόντιζε να μου το ξεκαθαρίσει ότι είναι μαλάκας. Κι εγώ το εκτίμησα δεόντως. Γιατί η ειλικρίνεια



Νέο



Το Drive μου



Σε κοινή χρήση



Πρόσφατα



Με αστερί



Κάδος απορ



Αποθηκευ

Χρησιμοποιώ
15 GBΑγορά απορ
κώρου

Νυχτερινό δελτίο -Πέτρος Μάρκαρης

Κάθε πρωί στις εννιά κοιταζόμαστε. Αυτός είναι όρθιος μπροστά στο γραφείο μου, με το βλέμμα κολλημένο πάνω μου, όχι ακριβώς στο ύψος των ματιών, αλλά λίγο πιο ψηλά, κάπου εκεί ανάμεσα στη βάση του μετώπου και στις βλεφαρίδες. «Είμαι μαλάκας» μου λέει.

Όχι με λόγια, με το βλέμμα του μου το λέει. Εγώ είμαι καθισμένος πίσω από το γραφείο και τον κοιτάζω ακριβώς στην κόρη, ούτε πιο ψηλά ούτε πιο χαμηλά. Γιατί εγώ είμαι ο προϊστάμενος του και μπορώ να τον κοιτάζω στα μάτια, αυτός δεν μπορεί. «Το ξέρω ότι είσαι μαλάκας» του λέω. Ούτε εμένα βγαίνει λέξη από το στόμα μου, το βλέμμα μου μιλάει. Αυτή τη συζήτηση την κάνουμε δέκα μήνες το χρόνο, αν εξαιρέσεις τους δύο που είμαστε ο καθένας σε άδεια, πέντε μέρες την εβδομάδα, Δευτέρα ως Παρασκευή, χωρίς να λέμε λέξη, μόνο με το βλέμμα. «Είμαι μαλάκας»- «Το ξέρω ότι είσαι μαλάκας».

Σε κάθε υπηρεσία αναλογεί κι ένα ποσοστό αποτυχημένων. Δεν μπορείς να 'χεις μόνο σαΐνια, θα φορτωθείς και μερικούς στούρνους. Ο Θανάσης ανήκει σ' αυτό το ποσοστό. Μπήκε στη Σχολή Αξιωματικών της Αστυνομίας, αλλά τα παράτησε στη μέση. Με τα χίλια βάσανα κατάφερε να φτάσει αρχιφύλακας, κι εκεί άραξε. Καμία φιλοδοξία για παραπάνω. Από την πρώτη μέρα που ήρθε στην υπηρεσία, φρόντισε να μου το ξεκαθαρίσει ότι είναι μαλάκας. Κι εγώ το εκτίμησα δεόντως. Γιατί η ειλικρίνειά του τον έσωσε από τις δύσκολες αποστολές, τα ξενύχτια, τα μπλόκα, τα κυνηγητά. Τον κράτησα στο γραφείο. Καμία εύκολη ανάκριση, αρχειοθέτηση, τις επαφές με την ιατροδικαστική υπηρεσία, το υπουργείο. Επειδή όμως έχουμε χρόνια έλλειψη ανδρών στο Σώμα, τρέχουμε και δε φτάνουμε, αυτός φροντίζει να μου το θυμίζει κάθε μέρα ότι είναι μαλάκας, μην το ξεχάσω και βρεθεί κατά λάθος σε κανένα

3ο ΣΤΑΔΙΟ: ΕΠΙΣΗΜΕΙΩΣΗ & ΛΗΜΜΑΤΟΠΟΙΗΣΗ ΤΟΥ CORPUS

- **Parser ILSP (PoS-tagged and lemmatized tokens with ILSP tools (Papageorgiou et al., 2000).**

```
40 . . PUNCT PTERM_P PunctType=PTERM_P 4 punct _ _  
  
# sent_id = Markaris_Nyxterino_Deltio313  
# text = Δεν ξέρει τίποτα, ψαρεύει στα θολά.  
1 Δεν δεν PART PtNg Particietype=Neg 2 advmod _ _  
2 ξέρει ξέρω VERB VbMn Aspect=Imp|Mood=Ind|Number=Sing|Person=3|  
3 τίποτα τίποτα PRON PnId Case=Acc|Gender=Neut|Number=Sing|Person=3  
4 , , PUNCT PUNCT _ 5 punct _ _  
5 ψαρεύει ψαρεύω VERB VbMn Aspect=Imp|Mood=Ind|Number=Sing|Person=3|  
6 στα του ADP AsPpPa Case=Acc|Gender=Neut|Number=Plur 7 case _  
7 θολά θολός ADJ Aj Case=Acc|Gender=Neut|Number=Plur 5 obl _ S  
8 . . PUNCT PTERM_P PunctType=PTERM_P 2 punct _ _  
  
# sent_id = Markaris_Nyxterino_Deltio314  
# text = «Λες να υπήρχαν παιδιά και να μην τα είδαμε;»  
1 « « PUNCT OPUNCT PunctType=OPUNCT 2 punct _ SpaceAfter=No
```

Μ' έχω φέρω ως ο λαιμός
αγριεύω ο βλέμμα μου
αγρός ηγόραζε
αδειάζω εγώ και ο δύο μου ο γωνιά
αδερφός να εγώ πετυχαίνω
ακονίζω ο μυαλό μου
ακούω ο εξάψαλμος
αλαφρύνεις ο θέση μου
αλλά εγώ πατάω
αλλάζω ακόμα ένας ματιά
αλλάζω δέκα χρώμα
αλλάζω ματιά
αλλάζω ο Μανωλιός
αλλάζω ο τροπάριο
αλλάζω ρότα
αλλάζω τροπάριο
αλλάζω χρώμα
άλλος Γκίκας βάζω στου κεφάλι μου
άμα εγώ πιάνω προκοπή
ανάβω και κορώνω
αναισθησία εγώ δέρνω
ανακατεύω τε άντερο μου
αν βάζω ο σκέψη μου στου χαρτί
αν δεν βγάζω άκρη
ανεβάζω ο ηθικός
ανεβάζω πυρετός

4ο ΣΤΑΔΙΟ: ΔΗΜΙΟΥΡΓΙΑ ΦΡΑΣΤΙΚΩΝ ΜΟΤΙΒΩΝ / Patterns

ψάχνω ψύλλους στ' άχυρα

↓ ↓ ↓ ↓

Vb + **No** + **AsPp** + **No**

(Pn)+(Vb)+**Vb**+(At)+(**No**)+(Ad)+**AsPp**+(Aj)+(At)+**No**

Την πάτησε

↓ ↓

Pn + **Vb**

(Pn)+(At)+(No)+(Pt)+**Pn**+(Vb)+**Vb**

Final Patterns
(Pn)+(Vb)+ Vb +(Ad)+(At)+(Aj)+(At)+ No +(Pn)+(Aj)
Vb +(At)+(No)+ Cj +(Pn)+(At)+(No)
Vb + Cj +(Pt)+(Pt)+ Vb
(Pn)+(Vb)+ Vb +(Cj)+(Ad)+(At)+(No)+ AsPp +(Ad)+(At)+ No +(At)+(No)
(Pn)+(Vb)+ Vb +(At)+(No)+(Ad)+ AsPp +(Aj)+(At)+ No
(Pn)+(Vb)+ Vb + Ad +(Ad)+(Pn)
Vb + AsPp +(At)+ No + AsPp +(At)+ No
(Pn)+(At)+(No)+(Pt)+ Pn +(Vb)+ Vb
(Pn)+ Pn +(Pt)+(Pt)+(Vb)+ Vb +(Pn)+ No +(Pt)+(Pn)+(Vb)+ Vb

The Multiword Expressions toolkit is a tool that aids in the automatic identification of multiword units such as idiomatic expressions in large text bases, independently of the language. (Carlos Ramisch, "A Generic Framework for Multiword Expressions Treatment: from Acquisition to Applications")

- Εργαλείο NLP
- Έχει ήδη χρησιμοποιηθεί στα νέα ελληνικά για την ανακάλυψη ονοματικών ΠΛΕ (Linardaki et al. 2010)

Ορισμένες από τις δυνατότητες του εργαλείου:

1. Εξάγει λεξικό MWEs από επισημειωμένο σώμα κειμένων
2. Δίνει την δυνατότητα μη αυτόματης αξιολόγησης των αποτελεσμάτων
3. Υπολογίζει **τα αποτελέσματα της συμφωνίας μεταξύ των annotators** και την ορθότητα κάθε καταχώρησης στο υποψήφιο λεξικό MWE σε αντιπαραβολή με το λεξικό αναφοράς **(GS)**

ΟΙ ΠΑΡΑΚΑΤΩ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΤΡΗΣΕΙΣ ΜΑΣ ΕΠΙΣΤΡΕΦΟΥΝ ΤΑ ΠΑΡΑΚΑΤΩ ΑΠΟΤΕΛΕΣΜΑΤΑ

- **Dice** - 3,000
- **LL** - 3,000
- **Mle** - 3,000
- **PMI** - 3,000
- **T-Score** - 3,000

Προέκυψαν **15,000** υποψήφιες VMWEs

6ο ΣΤΑΔΙΟ: ΔΙΑΔΙΚΑΣΙΑ ΑΞΙΟΛΟΓΗΣΗΣ

- Αξιολογήσαμε 5 λίστες πιθανών πολυλεκτικών (η κάθε λίστα προήλθε από μία διαφορετική στατιστική μέτρηση).

Dice: 2grams, 3grams, 4grams και 5grams (έφερε αποτελέσματα που δεν έφεραν τα άλλα στατιστικά + άπαξ λεγόμενα)

LL: 2grams, 3grams και 4grams

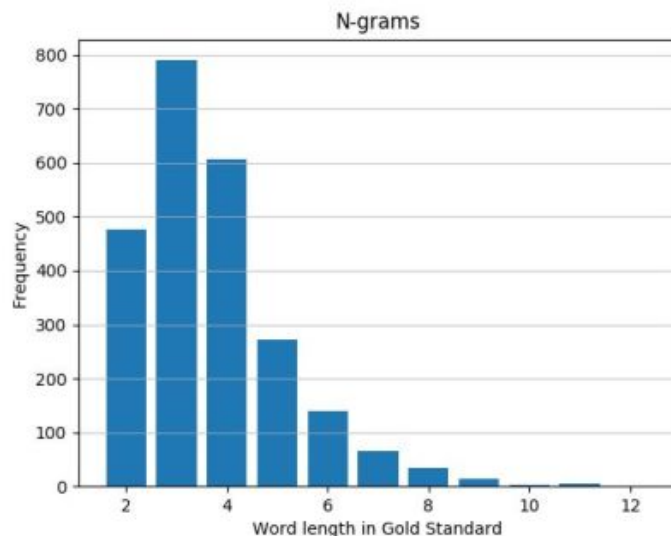
MLE: κυρίως 2grams, 4grams

T-score: κυρίως 2grams, 3grams

PMI: κυρίως 5grams

(έφερε αποτελέσματα που δεν έφεραν τα άλλα στατιστικά + άπαξ λεγόμενα)

*Τα αποτελέσματα αξιολογήθηκαν τόσο χειροκίνητα όσο και αυτόματα



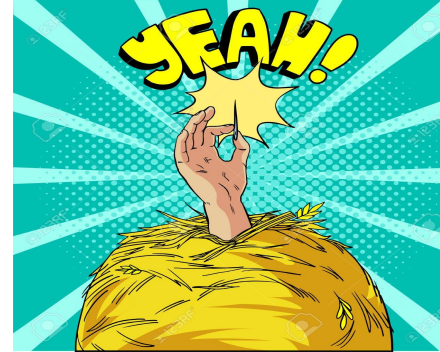
➤ Η δουλειά των annotators ήταν να ταξινομήσουν τα αποτελέσματα σε πολυλεκτικές ή μη.

*Ήταν σημαντικό κάθε annotator να μην επηρεάζεται από την άποψη του άλλου.

participle	wrong OCR	wrong tag
Association Measures		
PMI_plain_patterns_ngrams.txt	TP	FP
<cand candid=13035> παραλάμβανε τα παιδιά έξω από την Καστοριά		<cand candid=13035> παραλάμβανε τα παιδιά έξω από την Καστοριά
<cand candid=19702> προχώρησε ακολουθώντας τον Δήμο για τη γέφυρα		<cand candid=19702> προχώρησε ακολουθώντας τον Δήμο για τη γέφυρα
<cand candid=9231> μου μπήκαν οι ψύλλοι στ' αφτιά	<cand candid=9231> μου μπήκαν οι ψύλλοι στ' αφτιά	
του μπήκαν οι ψύλλοι στ' αφτιά	του μπήκαν οι ψύλλοι στ' αφτιά	
<cand candid=14853> είχε ξεπαρθενέψει τουλάχιστον το μισό χωριό		<cand candid=14853> είχε ξεπαρθενέψει τουλάχιστον το μισό χωριό
<cand candid=5990> κάνει μια έρευνα σχετικά με τον		<cand candid=5990> κάνει μια έρευνα σχετικά με
PMI_Annot_A/Pen	PMI_Annot_B/Art	LL_Annot_A/Mar 3 LL_Annot

ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ

- Αν διατηρείται η σημασία των MWEs σε περίπτωση μετάφρασης πχ στα αγγλικά ή και στα γαλλικά (Constant et al. 2017; Salehi et al. 2018)
 - Παίρνω τηλέφωνο – give a call – donner un coup de fil
 - Πιάνω τον ταύρο από τα κέρατα – take the bull by the horns – prendre le taureau par les cornes
 - Ψάχνω φύλλους στ' άχυρα – look for a needle in a haystack – chercher une aiguille dans une meule de foin



ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ

- Διάκριση μεταξύ VMWEs και κυριολεκτικών προτάσεων
πχ. πάω περίπατο

B	C	D
participle	wrong OCR	wrong tag
Association Measures		
Dice_plain_candidates.txt	TP	FP
<cand candid=28574>		<cand candid=28574>
ό,τι συμβαίνει		ό,τι συμβαίνει
ό,τι συνέβη		ό,τι συνέβη
<cand candid=29181>		<cand candid=29181>
δόσιμο νομίζοντας		δόσιμο νομίζοντας
<cand candid=372>	<cand candid=372>	<cand candid=372>
πάει περίπατο	πάει περίπατο (x3)	πάει περίπατο (x1)
πήγαινε περίπατο	πήγαινε περίπατο (x1)	πήγαινε περίπατο (x2)
πήγαν περίπατο		πήγαν περίπατο
πηγαίνουμε περίπατο		πηγαίνουμε περίπατο
<cand candid=4389>		<cand candid=4389>
που 'ξερα		που 'ξερα
που 'ξερες		που 'ξερες
που ήξερα		που ήξερα
που ήξεραν		που ήξεραν

1^ο ΠΕΙΡΑΜΑ

ΑΠΟΤΕΛΕΣΜΑΤΑ ΙΑ (Fleiss' kappa)

	A	B	C	D	E
1	Dice	MLE	LL	PMI	T
2	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters
3	Ni = 504 items	Ni = 1177 items	Ni = 528 items	Ni = 282 items	Ni = 1197 items
4	Nk = 498 categories	Nk = 1165 categories	Nk = 515 categories	Nk = 284 categories	Nk = 1137 categories
5	Nc x Ni = 1008 judgements	Nc x Ni = 2354 judgements	Nc x Ni = 1052 judgements	Nc x Ni = 564 judgements	Nc x Ni = 2394 judgements
6					
7	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:
8	multi-ao = 0.601190	multi-ao = 0.523364	multi-ao = 0.545627	multi-ao = 0.347518	multi-ao = 0.568087
9	multi-pi (Fleiss' kappa) = 0.583715	multi-pi (Fleiss' kappa) = 0.495323	multi-pi (Fleiss' kappa) = 0.519783	multi-pi (Fleiss' kappa) = 0.268323	multi-pi (Fleiss' kappa) = 0.546821
10	multi-kappa = 0.600700	multi-kappa = 0.523441	multi-kappa = 0.545104	multi-kappa = 0.346700	multi-kappa = 0.567049
11					
12	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:
13	alpha = 1.000000	alpha = 0.984060	alpha = 1.000000	alpha = 0.984506	alpha = 0.994578
14	alpha-kappa = 1.000000	alpha-kappa = 0.984023	alpha-kappa = 1.000000	alpha-kappa = 0.983036	alpha-kappa = 0.994531
15					
16					
17					
18		COMPARE WITH AUTOMATIC EVALUATION RESULTS (Gold Standard)			
19	Ran also with R package (irr) got exact same results	Scores ranked: (most reliable)	Scores ranked accord to GS evaluation:		
20		Dice	T-score		
21	kappam.fleiss(literature)	T-score	MLE		
22	Fleiss' Kappa for m Raters	LL	LL		
23		MLE	Dice		
24	Subjects = 503	Pmi	Pmi		
25	Raters = 2				
26	Kappa = 0.581				
27					

1° ΠΕΙΡΑΜΑ

ΑΠΟΤΕΛΕΣΜΑΤΑ IA WITH 3rd ANNOTATOR (Fleiss' kappa)

	A	B	C	D	E	F	G
1	T-score	PMI	Log Likelihood	Mle	Dice		
2							
3	Nc = 3 raters	Nc = 3 raters	Nc = 3 raters	Nc = 3 raters	Nc = 3 raters		
4	Ni = 3684 items	Ni = 282 items	Ni = 526 items	Ni = 1177 items	Ni = 504 items		
5	Nk = 915 categories	Nk = 285 categories	Nk = 515 categories	Nk = 1185 categories	Nk = 498 categories		
6	Nc x Ni = 11052 judgements	Nc x Ni = 846 judgements	Nc x Ni = 1578 judgements	Nc x Ni = 3531 judgements	Nc x Ni = 1512 judgements		
7	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:		
8	multi-ao = 0.624955	multi-ao = 0.582848	multi-ao = 0.697085	multi-ao = 0.682243	multi-ao = 0.734127		
9	multi-pi (Fleiss' kappa) = 0.605541	multi-pi (Fleiss' kappa) = 0.525700	multi-pi (Fleiss' kappa) = 0.688674	multi-pi (Fleiss' kappa) = 0.668913	multi-pi (Fleiss' kappa) = 0.726103		
10	multi-kappa = 0.624714	multi-kappa = 0.561800	multi-kappa = 0.696057	multi-kappa = 0.682000	multi-kappa = 0.733740		
11	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:		
12	alpha = 0.995085	alpha = 0.986499	alpha = 1.000000	alpha = 0.990294	alpha = 1.000000		
13	alpha-kappa = 0.995033	alpha-kappa = 0.986079	alpha-kappa = 1.000000	alpha-kappa = 0.990286	alpha-kappa = 1.000000		
14							
15							
16							
17							
18							
19	IA calculated including 3rd annotator	Scores ranked according to IA (manual evaluation)					
20							
21		Dice					
22		Log Likelihood					
23		Mle					
24		T-score					
25		Pmi					
26							

2° ΠΕΙΡΑΜΑ

ΑΠΟΤΕΛΕΣΜΑΤΑ ΙΑ (Fleiss' kappa)

1	Interannotator Agreement Dice	Interannotator Agreement LL	Interannotator Agreement MLE	Interannotator Agreement PMI	Interannotation Agreement T-score	
2	Nc = 2 raters		Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	
3	Ni = 481 items	Nc = 2 raters	Ni = 480 items	Ni = 388 items	Ni = 439 items	
4	Nk = 487 categories	Ni = 311 items	Nk = 480 categories	Nk = 388 categories	Nk = 441 categories	
5	Nc x Ni = 962 judgements	Nk = 311 categories	Nc x Ni = 960 judgements	Nc x Ni = 776 judgements	Nc x Ni = 878 judgements	
6		Nc x Ni = 622 judgements				
7	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	
8	multi-ao = 0.858628	multi-ao = 0.729904	multi-ao = 0.852083	multi-ao = 0.842784	multi-ao = 0.833713	
9	multi-pi (Fleiss' kappa) = 0.857753	multi-pi (Fleiss' kappa) = 0.723921	multi-pi (Fleiss' kappa) = 0.850946	multi-pi (Fleiss' kappa) = 0.841	multi-pi (Fleiss' kappa) = 0.832219	
10	multi-kappa = 0.858375	multi-kappa = 0.729271	multi-kappa = 0.851821	multi-kappa = 0.842441	multi-kappa = 0.833397	
11						
12	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	
13	alpha = 0.985733	alpha = 1.000000	alpha = 1.000000	alpha = 1.000000	alpha = 0.994601	
14	alpha-kappa = 0.985716	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 0.994595	
15						
16						
17						
18						
19						
20						
21						
22						

ΑΠΟΤΕΛΕΣΜΑΤΑ

= 6.654 VMWEs

Subcorpora	Tokens	Lemmas	Number of sentences	Number of MWEs
Maratos (2007)	84,172	5,937	5,931	1,136
Markaris (1995)	105,575	5,916	8,457	1,877
Markaris (2016)	84,172	4,86	5,931	1,175
Papadaki(2001)	82, 084	6,281	10,311	912
Tachtsis (1970)	104,215	6,701	6,700	1,554

7ο ΣΤΑΔΙΟ : ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ - ΛΙΣΤΕΣ

mwetoolkit3

- Το εργαλείο δεν επέστρεφε μόνο πολυλεκτικά
- Τα patterns μας επέστρεφαν πολλές φορές την ίδια φράση (1ο πείραμα) 
- Wrong OCR
- Wrong tag

Βελτιώσαμε τα Patterns και επαναλάβαμε την διαδικασία εκ νέου (2ο πείραμα)

Wrong OCR

Association Measures			Μετοχές
LL_plain_patterns_ngrams.txt	TP	FP	Wrong OCR
τραγουδώ_συνεχώς		τραγουδώ_συνεχώς	
κάνω_γούστο	κάνω_γούστο		
Λει_καράβι		Λει_καράβι	
μεταφέρω_ακριβώς		μεταφέρω_ακριβώς	
περιμένω_μπροστά		περιμένω_μπροστά	
εγώ_συστήνω		εγώ_συστήνω	
εγώ_ψιθυρίζω		εγώ_ψιθυρίζω	
εγώ_ξέρω		εγώ_ξέρω	
ξυπνώ_εύκολα		ξυπνώ_εύκολα	
υπηρετώ_πιστά		υπηρετώ_πιστά	
σχεδιάζω_αποβραδīs		σχεδιάζω_αποβραδīs	
διναν_κλωτσιά		διναν_κλωτσιά	
παίρνω_όρκος	παίρνω_όρκος		
συμφωνώ_σιωπηλά		συμφωνώ_σιωπηλά	

Wrong TAG

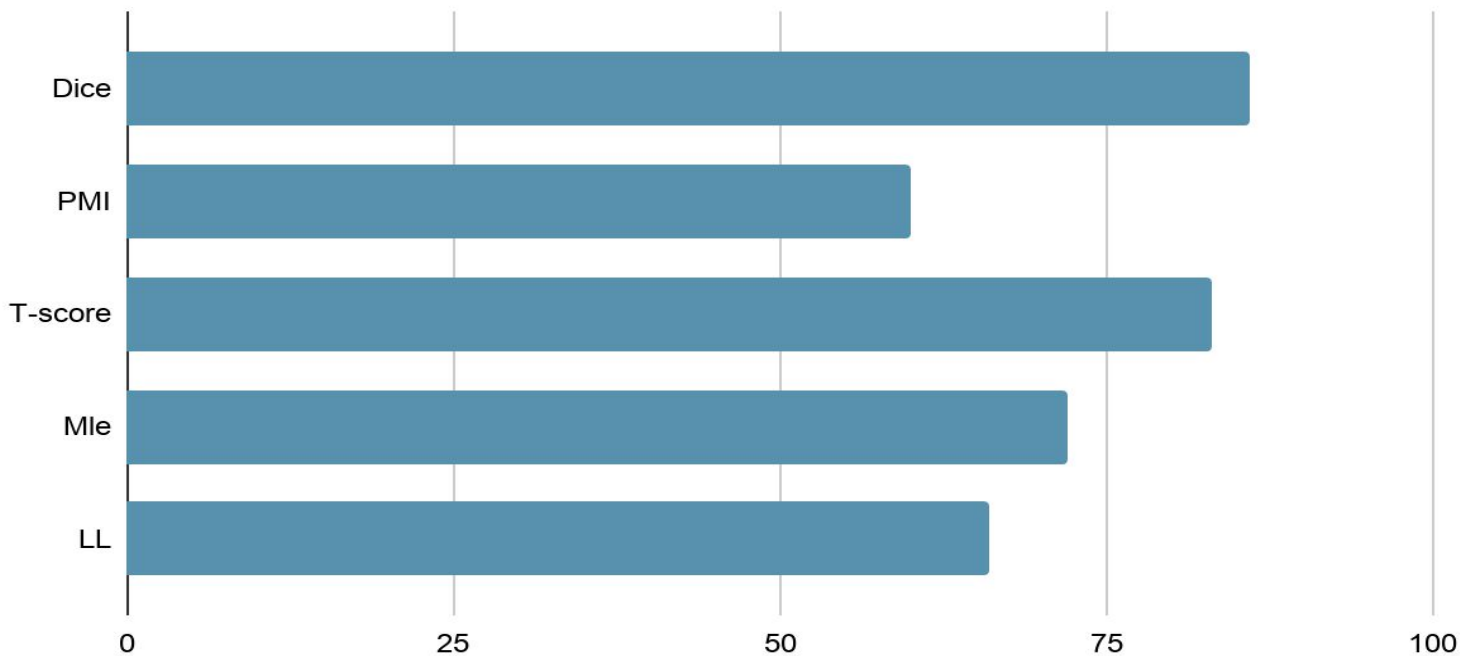
Dice_plain_candidates.txt	TP	FP
Λέξη γέμια		Λέξη γέμια
<cand candid=39327>		<cand candid=39327>
φορούσαν στρατιωτικά χιτάνια		φορούσαν στρατιωτικά χιτάνια
<cand candid=1903>	<cand candid=1903>	
χαμόγελο μοραίνεται	χαμόγελο μοραίνεται	
<cand candid=21746>		<cand candid=21746>
χρειαζόταν διερμηνεία		χρειαζόταν διερμηνεία
<cand candid=23635>		<cand candid=23635>
οποία μοιράστηκαν		οποία μοιράστηκαν
<cand candid=30616>		<cand candid=30616>
χρονώ κορίτσι		χρονώ κορίτσι
<cand candid=31084>		<cand candid=31084>
Θόδωρος παράτησε		Θόδωρος παράτησε
<cand candid=48074>	<cand candid=48074>	
τραβάρ ρουφηξιά	τραβάρ ρουφηξιά	
<cand candid=10353>		<cand candid=10353>
φτιάχνει πολιτικό παγωτό		φτιάχνει πολιτικό παγωτό
<cand candid=22057>		<cand candid=22057>
Μπερδεμένα πράγματα		Μπερδεμένα πράγματα
<cand candid=36926>		<cand candid=36926>
διαβάζει Ντελλό		διαβάζει Ντελλό
<cand candid=41351>		<cand candid=41351>
διαβάζει ρομάντζο		διαβάζει ρομάντζο
<cand candid=42951>		<cand candid=42951>

7ο ΣΤΑΔΙΟ : ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ - ΛΙΣΤΕΣ mwetoolkit3

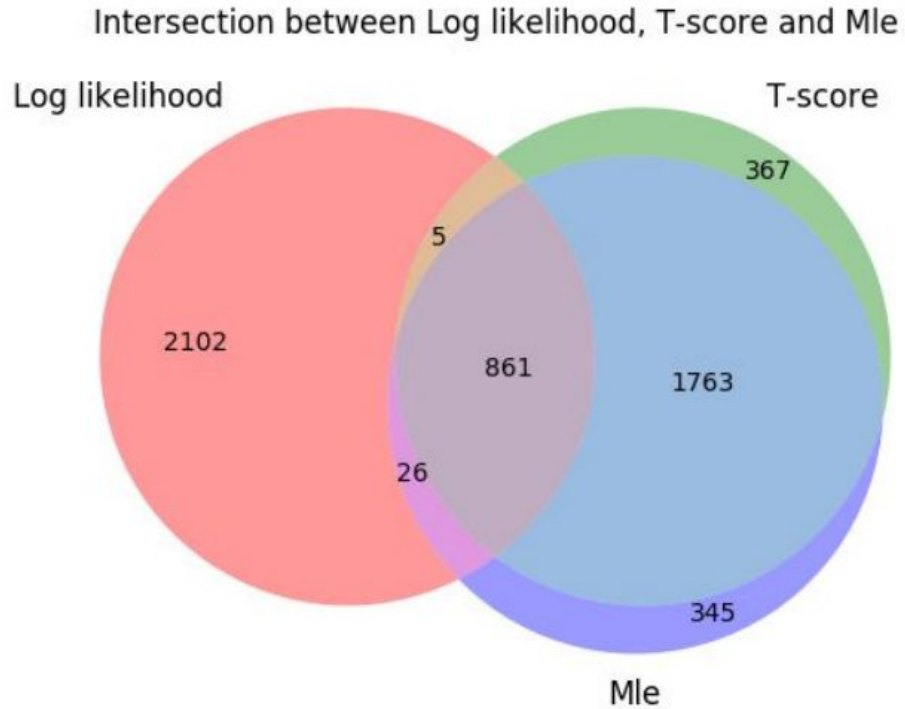
- Τα εμπλουτισμένα φραστικά μοτίβα βελτίωσαν τα αποτελέσματα μας (Stamou et al. 2020)
- Η δουλειά των annotators ήταν σημαντική καθώς εντόπισαν και πιο ασυνήθιστες VMWEs
- Το Dice αποδείχθηκε το πιο αξιόπιστο AM και στα δύο πειράματα
- Το PMI επέστρεψε VMWEs που δεν βρέθηκαν από τα άλλα AM

ΑΠΟΤΕΛΕΣΜΑΤΑ Fiction Corpora

Ποσοτική Αξιολόγηση



Ποιοτική Αξιολόγηση



Twitter corpora



ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ

❖ Αφαιρέσαμε:

1. mentions
2. https links
3. hashtags
4. emoticons

❖ ILSP Parser



```
19 # text = Αυτος ειναι ορθιος μπροστα στο γραφειο μου, με το βλεμμα κολλημενο πανω μου, οχι ακριβως στο
20 1 Αυτός αυτός PRON PnDm Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Dem 3 nsubj
21 2 είναι είμαι VERB VbMn Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|Voice=Pass 3
22 3 όρθιος όρθιος ADJ Aj Case=Nom|Gender=Masc|Number=Sing 0 root _ _
23 4 μπροστά μπροστά ADV Ad _ 3 advmod _ _
24 5 στο του ADP AsPpPa Case=Acc|Gender=Neut|Number=Sing 6 case _ _
25 6 γραφείο γραφείο NOUN NoCm Case=Acc|Gender=Neut|Number=Sing 4 obl _ _
26 7 μου μου PRON PnPp Case=Gen|Gender=Masc|Number=Sing|Person=1|Poss=Yes|PronType=Prs 6 nmod
27 8 , , PUNCT PUNCT _ 11 punct _ _
28 9 με με ADP AsPpSp _ 11 case _ _
29 10 το ο DET AtDf Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art 11 det _ _
30 11 βλέμμα βλέμμα NOUN NoCm Case=Acc|Gender=Neut|Number=Sing 3 obl _ _
31 12 κολλημένο κολλημένος ADJ Aj Case=Acc|Gender=Masc|Number=Sing 11 amod _ _
32 13 πάνω πάνω ADV Ad _ 12 advmod _ _
33 14 μου εγώ PRON PnPp Case=Gen|Gender=Masc|Number=Sing|Person=1|PersPronType=We|PronType=Prs 13
34 15 , , PUNCT PUNCT _ 19 punct _ _
35 16 όχι όχι PART PtNg ParticleType=Neg 17 advmod _ _
36 17 ακριβώς ακριβώς ADV Ad _ 19 advmod _ _
37 18 στο του ADP AsPpPa Case=Acc|Gender=Neut|Number=Sing 19 case _ _
38 19 ύψος ύψος NOUN NoCm Case=Acc|Gender=Neut|Number=Sing 11 conj _ _
39 20 των ο DET AtDf Case=Gen|Definite=Def|Gender=Neut|Number=Plur|PronType=Art 21 det _ _
40 21 ματιών μάτι NOUN NoCm Case=Gen|Gender=Neut|Number=Plur 19 nmod _ SpaceAfter=No
41 22 , , PUNCT PUNCT _ 26 punct _ _
```

ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ Twitter

1ο ΠΕΙΡΑΜΑ- Baseline

- Χρησιμοποιήσαμε τα προηγούμενα φραστικά μοτίβα
- Αξιολογήσαμε 3.000 λημματοποιημένες φράσεις/Στατιστικό
- Σειρά κατάταξης σύμφωνα με το ΙΑ:

1. MLE
2. T-score
3. LL
4. Dice
5. PMI

- Σειρά κατάταξης σύμφωνα με τον αριθμό των φράσεων:

1. PMI
2. LL
3. MLE
4. T-score
5. Dice

1ο Πείραμα – Αποτελέσματα ΙΑ (Fleiss' kappa)

	A	B	C	D	E	F
1	Dice	Log likelihood	Mle	Pmi	T-score	
2	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	
3	Ni = 105 items	Ni = 148 items	Ni = 89 items	Ni = 202 items	Ni = 93 items	
4	Nk = 96 categories	Nk = 149 categories	Nk = 89 categories	Nk = 202 categories	Nk = 93 categories	
5	Nc x Ni = 210 judgements	Nc x Ni = 296 judgements	Nc x Ni = 178 judgements	Nc x Ni = 404 judgements	Nc x Ni = 186 judgements	
6						
7	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	Overall agreement coefficients for all annotators:	
8	multi-ao = 0.628571	multi-ao = 0.716216	multi-ao = 0.831461	multi-ao = 0.534653	multi-ao = 0.795699	
9	multi-pi (Fleiss' kappa) = 0.610519	multi-pi (Fleiss' kappa) = 0.708525	multi-pi (Fleiss' kappa) = 0.828230	multi-pi (Fleiss' kappa) = 0.505128	multi-pi (Fleiss' kappa) = 0.791159	
10	multi-kappa = 0.625857	multi-kappa = 0.714849	multi-kappa = 0.829871	multi-kappa = 0.533430	multi-kappa = 0.793936	
11						
12	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	Weighted agreement coefficients for all annotators:	
13	alpha = 1.000000	alpha = 0.990789	alpha = 1.000000	alpha = 1.000000	alpha = 1.000000	
14	alpha-kappa = 1.000000	alpha-kappa = 0.990717	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 1.000000	
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						

ΑΞΙΟΛΟΓΗΣΗ Twitter

2ο ΠΕΙΡΑΜΑ

- Βελτιώσαμε τα φραστικά μοτίβα (μέχρι 4 μέρη του λόγου)
- Επαναλάβαμε την ίδια διαδικασία

Γιατί προέκυψε η ανάγκη για 2ο πείραμα:

Πλούσια Φραστικά Μοτίβα που επέστρεφαν τις ρηματικές ΠΛΕ με πολύ περικείμενο.



2ο Πείραμα – Αποτελέσματα ΙΑ (Fleiss' kappa)

	A	B	C	D	E
1	Dice	Log likelihood	MLE	PMI	T-score
2					
3	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters	Nc = 2 raters
4	Ni = 98 items	Ni = 93 items	Ni = 88 items	Ni = 89 items	Ni = 108 items
5	Nk = 73 categories	Nk = 38 categories	Nk = 48 categories	Nk = 80 categories	Nk = 61 categories
6	Nc x Ni = 196 judgements	Nc x Ni = 186 judgements	Nc x Ni = 176 judgements	Nc x Ni = 178 judgements	Nc x Ni = 216 judgements
7					
8	Overall agreement coefficients for all annotators: multi-ao = 0.877551	Overall agreement coefficients for all annotators: multi-ao = 0.709677	Overall agreement coefficients for all annotators: multi-ao = 0.818182	Overall agreement coefficients for all annotators: multi-ao = 0.685393	Overall agreement coefficients for all annotators: multi-ao = 0.768519
10	multi-pi (Fleiss' kappa) = 0.874714	multi-pi (Fleiss' kappa) = 0.691125	multi-pi (Fleiss' kappa) = 0.809472	multi-pi (Fleiss' kappa) = 0.671803	multi-pi (Fleiss' kappa) = 0.759154
11	multi-kappa = 0.875397	multi-kappa = 0.700537	multi-kappa = 0.812592	multi-kappa = 0.681534	multi-kappa = 0.764706
12					
13	Weighted agreement coefficients for all annotators: alpha = 1.000000	Weighted agreement coefficients for all annotators: alpha = 1.000000	Weighted agreement coefficients for all annotators: alpha = 1.000000	Weighted agreement coefficients for all annotators: alpha = 1.000000	Weighted agreement coefficients for all annotators: alpha = 1.000000
14	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 1.000000	alpha-kappa = 1.000000
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					

Οι στατιστικές μετρήσεις Mle, T-Score, LL παρουσιάζουν πολύ μεγάλη επικάλυψη(πολλά κοινά στοιχεία).

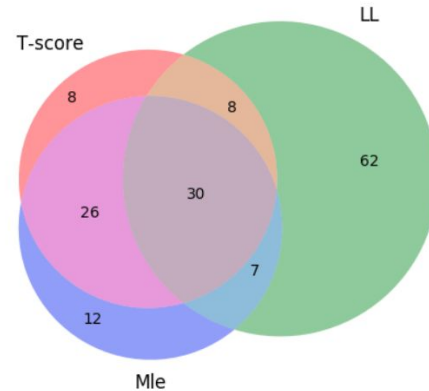
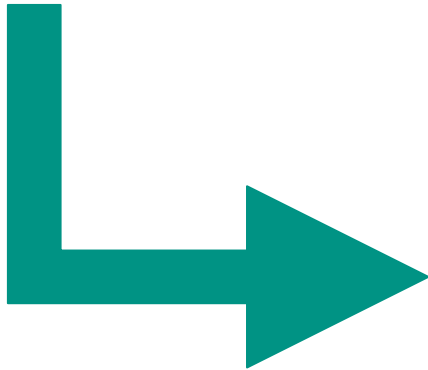


Figure 1: Intersection among T-score, LL and MLE True positives.

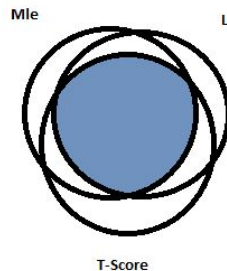
ΑΞΙΟΛΟΓΗΣΗ Twitter

3ο ΠΕΙΡΑΜΑ

Γιατί προέκυψε η ανάγκη για 3ο πείραμα:

Μεγάλη επικάλυψη των ΣΤΑΤΙΣΤΙΚΩΝ μετρήσεων MLE, LL και T.

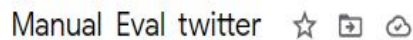
- αξιολόγηση της τομής των στατιστικών μετρήσεων MLE, LL και T
- Το DICE και το PMI αξιολογήθηκαν ξεχωριστά
- Αυξήσαμε τον αριθμό των υποψηφίων φράσεων (+3000)



Τι συναντήσαμε στις λίστες Twitter κατά την αξιολόγηση:

- Greeklish και Engreek (πχ. γουι_λοβ)
- Κυπριακά (πχ. τζαι_αναρωθκιουμαστε_ούλος_αδελφός)
- Αργκό (πχ. ρίχνω_πιστόλι)
- Πολύ ευφάνταστες ρηματικές ΠΛΕ (πχ. κλάνω_μέντουλες)

***Τα κριτήρια αξιολόγησης ήταν τα ίδια με την προηγούμενη φορά



[Αρχείο](#)
[Επεξεργασία](#)
[Προβολή](#)
[Εισαγωγή](#)
[Μορφή](#)
[Δεδομένα](#)
[Εργαλεία](#)
[Πρόσθετα](#)
[Βοήθεια](#)
[Η τελευταία τροποποίηση πραγ...](#)

 Κοινή Χρήση

P



<i>fx</i>	Color meaning
-----------	---------------

	A	B	C	D
1	Color meaning	participle	wrong OCR	wrong tag
2	VMwe candidates (5000 highest-ranking) patterns & ngrams	Association Measures		
3		# MWETOOLKIT: filetype="PlainCandidates"	TP	FP
4		αυτός_τούς_ψηφίζω_ο_τροπολογία_για_ο_διευκόλυνση_ο_δωρεά		αυτός_τούς_ψηφίζω_ο_τροπολογία_α_ο_διευκόλυνση_ο_δωρεά
5		παραμυθάς_ρίχνω_ο_φταίξιμο_στον_αριστερά_έναν_ζωή_ληστοσυμμοριτής	παραμυθάς_ρίχνω_ο_φταίξιμο_στον_αριστερά_έναν_ζωή_ληστοσυμμοριτής	
6		τούς_ψηφίζω_ο_τροπολογία_για_ο_διευκόλυνση_ο_δωρεά		τούς_ψηφίζω_ο_τροπολογία_για_ο_ευκόλυνση_ο_δωρεά
7		εγώ_γιορτάζω_και_φέτος_ο_χάρτης_μ_ε_ο_καταυλισμός_ο_επενδυτής		εγώ_γιορτάζω_και_φέτος_ο_χάρτης_ε_ο_καταυλισμός_ο_επενδυτής
8		εγώ_συνδιοργανώνω_τελώ_ο_υπό_ο_αιγίδα_ο_Δήμος	εγώ_συνδιοργανώνω_τελώ_ο_υπό_ο_αιγίδα_ο_Δήμος	
9		που_παρακάμπτω_και_ο_μπέρδεμα_με_ο_χρόνος_ο_ρήμα		που_παρακάμπτω_και_ο_μπέρδεμα_μ_ο_χρόνος_ο_ρήμα



 ar

MLE Annot A/Penny ▾

MLE Annot B/Mar ▾

PMI Annot A/Art ▼

PMI Annot B/Mar ▼

Εξερευνήστε



ΣΥΜΠΕΡΑΣΜΑΤΑ

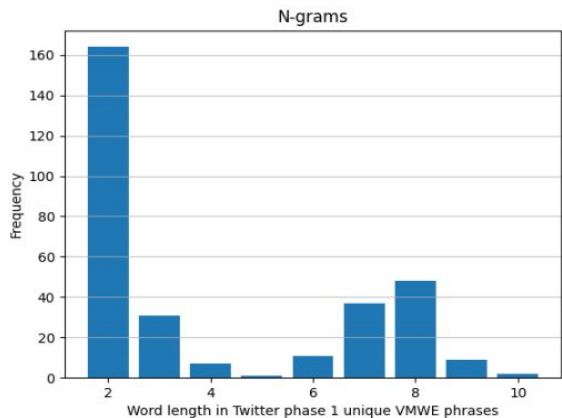
- Η τομή των LL-MLE-T-score παρουσιάζουν υψηλή συμφωνία (IA) σε αντίθεση με το PMI και το Dice
- Σειρά κατάταξης σύμφωνα με τον αριθμό των φράσεων:
 1. PMI
 2. Τομή (LL, MLE, T-score)
 3. Dice
- 431 Νέες φράσεις



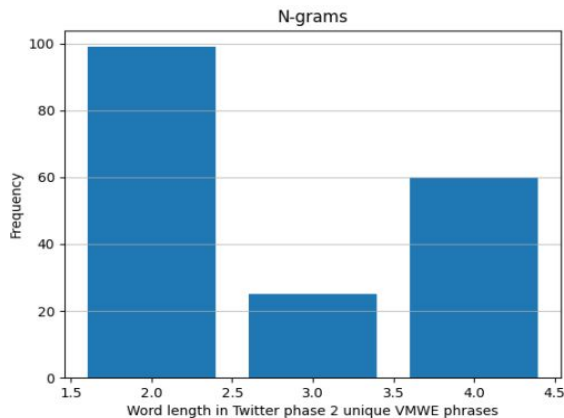
ΣΥΜΠΕΡΑΣΜΑΤΑ

- Η χειροκίνητη αξιολόγηση δεν μπορεί να αποφευχθεί
- Πιο αποτελεσματικό:
αξιολόγηση του Dice και του PMI ανεξάρτητα και την τομή LL, MLE και T-score

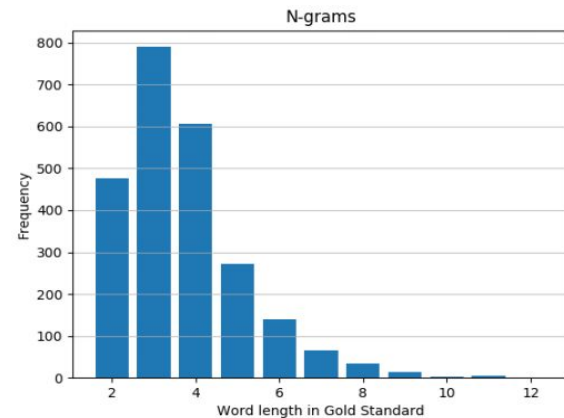
ΣΥΜΠΕΡΑΣΜΑΤΑ



(a) Twitter-Baseline



(b) Twitter-lean patterns



(c) Fiction

ΣΥΓΚΡΙΣΗ LITERATURE CORPORA & TWITTER

- ❖ Στα λογοτεχνικά σώματα κειμένων έχουμε πλούσια συντακτικά μοτίβα
- ❖ Στα λογοτεχνικά σώματα κειμένων έχουμε πιο προσεγμένο λόγο σε αντίθεση με τα tweets που είχαν πιο προφορικό λόγο
- ❖ Οι επισημειωτές εντόπισαν λιγότερο συνηθισμένες VMWEs σε αντίθεση με την αυτόματη αξιολόγηση (Mwetoolkit3)
- ❖ Το DICE ήταν το πιο αξιόπιστο σκορ ενώ το PMI επέστρεψε τα άπαξ λεγόμενα

"Πιάσ' το αβγό και κούρευ' το"

Έγινεεεεε!!!



Βιβλιογραφία

[Maratos (2007)] Μαράτος, Τ. (2007). Οι τυφώνες ήταν γένους θηλυκού. Βιβλιοπωλείον της “Εστίας”, Ι.Δ. ΚΟΛΛΑΡΟΥ & ΣΙΑΣ.

[Markaris (1995)] Μάρκαρης, Π. (1995). Νυχτερινό Δελτίο, ΓΑΒΡΙΗΛΙΔΗΣ.

[Markaris (2016)] Μάρκαρης, Π. (2016). Offshore. ΓΑΒΡΙΗΛΙΔΗΣ.

[Papadaki (2001)] Παπαδάκη, Α. (2001). Βαρκάρισσα της χίμαιρας. Καλέντης.

[Tachtsis (1970)] Ταχτσής, Κ. (1970). Το τρίτο στεφάνι. ΕΡΜΗΣ.

Vivian Stamou, Marilena Malli, Penny Takorou, Artemis Xylogianni, Stella Markantonatou. 2020. Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek. *Proceedings of the Euralex XIX: Lexicography for Inclusion*. Alexandroupolis, Greece (to appear).

Vivian Stamou, Marilena Malli, Penny Takorou, Artemis Xylogianni, Stella Markantonatou. 2020. VMWE discovery: a comparative analysis between Literature and Twitter Corpora. COLING-2020 Proceedings

Για οποιαδήποτε ερώτηση/πληροφορία μη διστάσετε να μας στείλετε μήνυμα στις παρακάτω διευθύνσεις:

(Άρτεμις) artemis.xylo@gmail.com

(Μαριλένα) mallimarilena@gmail.com

(Πέννυ) pennytak07@gmail.com

(Σοφία) sophia.katsiouris24@gmail.com

MERCI DE VOTRE ATTENTION

