ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

«ΑΠΟΛΛΩΝΙΣ»

Γλωσσικοί πόροι στο ΕΚΠΑ: Υποδομή, χρήση και εργαλεία

Ειρήνη Φλώρου

ΕΚΠΑ, 01/03/2019

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στόχοι

- Μετατροπή ανεπεξέργαστων δεδομένων σε χρήσιμη πληροφορία χωρίς εξαντλητική επισημείωση
- Συνόψιση πληροφορίας κειμένου σε διαγράμματα με σημασία
- Αναπαράσταση της γνώσης με πολλαπλούς τρόπους
- Πλήρης κατανόηση
- Συγκράτηση πληροφορίας από τον χρήστη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Δυνατότητες

1. Εξόρυξη πληροφορίας
2. Σύγκριση δεδομένων
3. Αναγνώριση προτύπων

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Κατηγορίες τεχνικών οπτικοποίησης δεδομένων

1. αποκάλυψη περιεχομένου εγγράφων
2. απεικόνιση ομοιότητας εγγράφων
3. οπτικοποίηση συναισθημάτων που διατυπώνονται στο κείμενο
4. μελέτη ενός corpus στο οποίο ανήκει ένα έγγραφο
5. Ανάλυση δεδομένων συγκεκριμένου corpus π.χ. από δεδομένα μέσων κοινωνικής δικτύωσης, ηλεκτρονικές ειδήσεις, ποίηση

Γλωσσικά δεδομένα του CLARIN:EL

- Α. Κειμενικής μορφής
- Β. Προφορικού λόγου (ηχογραφήσεις)

Από ένα εύρος επιστημονικών πεδίων: νομικά σώματα κειμένων, ιατρικά λεξικά, οικονομικά και πολιτικά κείμενα, κείμενα σχετικά με την επιστήμη των υπολογιστών, τις γεωργικές επιστήμες, τη γεωλογία κ.ά.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Γλωσσικοί πόροι του CLARIN:EL

491 ψηφιακοί γλωσσικοί πόροι

- 400 σώματα κειμένων
- 35 εργαλεία επεξεργασίας
- 46 λεξικά

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Υπηρεσίες γλωσσικής επεξεργασίας του CLARIN:EL

1. Λεκτικής ανάλυσης
2. Αναγνώρισης λέξεων, ορίων προτάσεων και μερών του λόγου
3. Μορφολογικής και συντακτικής ανάλυσης
4. Αναγνώρισης ονοματικών οντοτήτων, ορολογίας, άποψης και συναισθήματος

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τεχνικές οπτικοποίησης των δεδομένων του CLARIN:EL

1. Συννεφόλεξο (word cloud / tag cloud)
2. Γράφημα στοιβάδας (sunburst)
3. Γράφημα t-sne

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Συννεφόλεξο (word cloud / tag cloud)

- Άμεση και σαφής απεικόνιση των λέξεων-κλειδιών ενός εγγράφου για την αναπαράσταση του περιεχομένου του.
- Οπτικοποίηση της συχνότητας εμφάνισης μιας λέξης

Στόχοι του συννεφόλεξου

- Γρήγορος εντοπισμός της κύριας ιδέας του εγγράφου
- Έλεγχος λέξεων που χρησιμοποιούνται περισσότερο σε κείμενα συγκεκριμένης θεματολογίας
- Συνόψιση και οπτική αναπαράσταση σημασιολογικών σχέσεων
- Ανακάλυψη προτύπων σε επίπεδο λέξης π.χ. επαναλήψεις, συνεμφανίσεις

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Χαρακτηριστικά συννεφόλεξου

1. Ευανάγνωστη γραμματοσειρά
2. Η ιδιαίτερα βαρύνουσα λέξη αποδίδεται στο κέντρο του σύννεφου
3. Ζωντανά χρώματα για την οπτικοποίηση σημαντικών λέξεων
4. Επιλογή μεγέθους του σύννεφου από τον χρήστη
5. Χρήση λίστας stop words του ΣΕΚ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Συννεφόλεξο αρχείου του OROSSIMO CORPUS – Medicine & Health

διαβήτη
γλυκόζης
ινσουλίνης
ινσουλίνη
τύπου
μπορεί
ασθενείς
αίματος
σακχαρώδη
κυττάρων
υπεργλυκαιμία
περιπτώσεις
παγκρέατος
υποδοχέων
παρουσία
αυτή
γλυκόζη
αιμοσφαιρίνη
αιμοσφαιρίνης
Έτσι
καθώς
μετά
κατά
αλλά
έχει
έχουν
όμως
κύτταρα
διαταραχή
περίπτωση
εμφανίζεται
ομάδα
χρόνια
γυναίκες
χωρίς
μορφή
παράγοντες
φυσιολογική
διαβήτης
αποτέλεσμα
δοκιμασία
πλάσματος
έκκριση
μόνο
υπογλυκαιμία
αυτά
λιπόν
δηλαδή
Τέλος
προκαλεί
κατάσταση
ασθενών
ανοχής
χορήγηση
αντίδραση
υπάρχει
άλλη
τύπο
δράση
σακχαρώδους
θεραπευτική
αδυναμία
πρόσληψη
παράδειγμα
παράγοντες
υποδοχείς
αντιγόνων
γονίδιο
απώλεια
ασθενούς
σύνθεση
αντισώματα
υπομονάδα
πρόκληψης
ανεπάρκεια
αυτές
ακόμα
χώρα
αντίθετα
έλεγχος
ενώ
διέγερση
παρατεταμένη
κληρονομική
γλυκαγόνη
γλυκοζουρίας
εμπλέκεται
ευαισθησία
ευαισθησίας
ένα
επίπεδα
αιτίας
δράσης
αυξημένη
υπογλυκαιμία
παράγοντας
λίπους
ορού
πρέπει
πλάσματος
σχετικά
ιστούς
ειδικά
Σ.Δ
Α1
ιστών
σήμα
ορισμένες
μάλιστα
εκτός
δακρύνεται
ινσουλινοεξαρτώμενο
υπογλυκαιμίας
αυτού
Σακχαρώδη
πρόσληψης
σύρα
νόσο
ωρίς
ωρίς
ωρίς

Συννεφόλεξο του ίδιου αρχείου με επιλογή μικρότερου μεγέθους

αποΛωνίς
ΕΠΑνΕΚ 2014-2020
ΕΣΠΑ 2014-2020
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
clarin:el

Γράφημα στοιβάδας (sunburst)

- Ακτινωτή και όχι ορθογώνια διάταξη των δεδομένων
- Η κορυφή της ιεραρχίας τοποθετείται στο κέντρο
- Τα βαθύτερα επίπεδα βρίσκονται μακρύτερα από το κέντρο

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στόχοι γραφήματος στοιβάδας

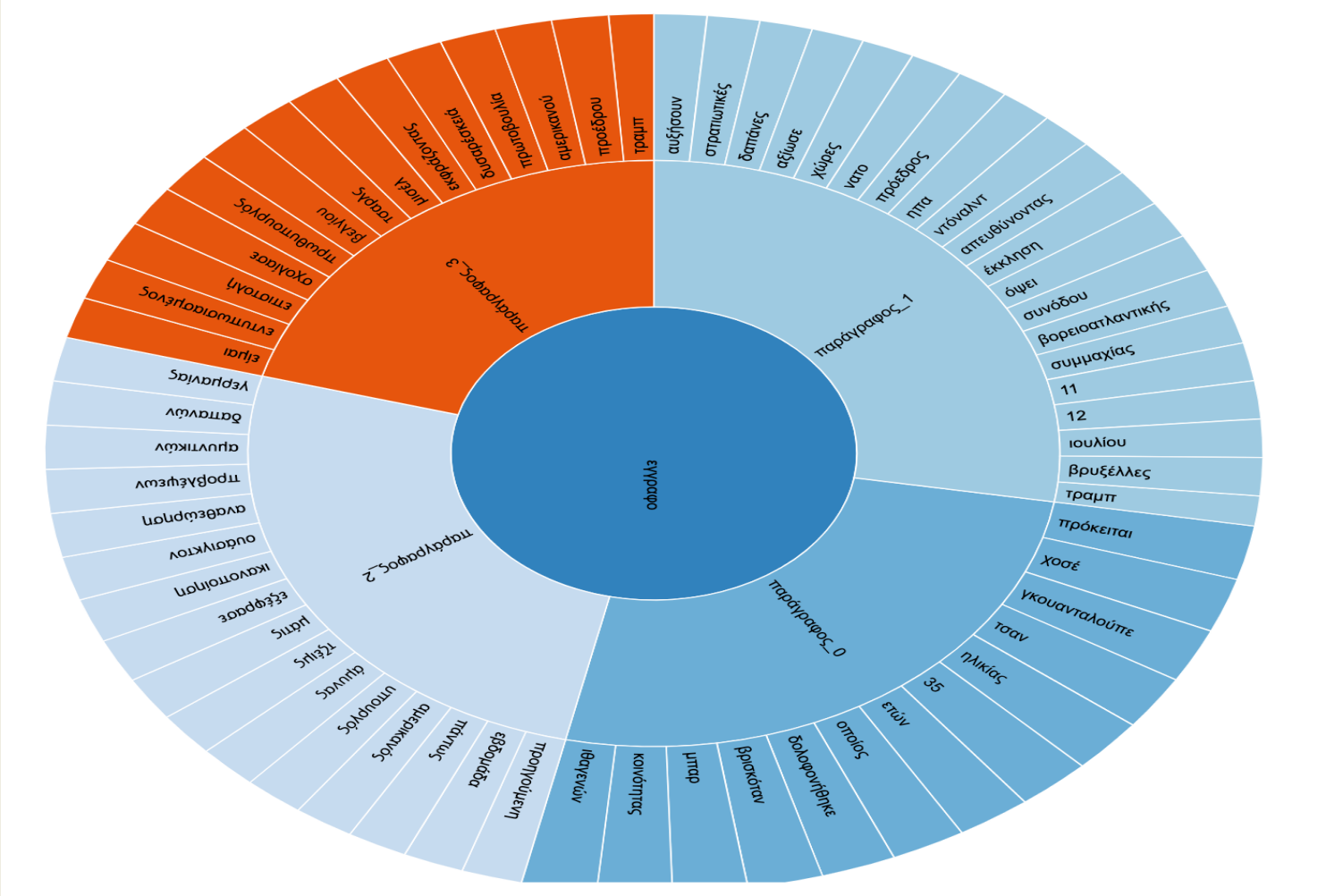
- Κατάδειξη του τρόπου με τον οποίο οι εξωτερικοί δακτύλιοι σχετίζονται με τους εσωτερικούς
- Κατάδειξη του τρόπου με τον οποίο ένας δακτύλιος είναι διαχωρισμένος στα επί μέρους τμήματά του

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Χαρακτηριστικά γραφήματος στοιβάδας

- Ακτινωτή απεικόνιση των δεδομένων
- Το κοινό χρώμα αποδίδει και κοινή ιδιότητα των δεδομένων
- Η γωνία κάθε φέτας είτε διαιρείται ισόποσα κάτω από τον γονικό της κόμβο είτε μπορεί να είναι ανάλογη με μια προκαθορισμένη τιμή
- Χρήση λίστας stop words του ΣΕΚ

Γράφημα στοιβάδας (sunburst)



από λόγος
ΕΠΑνΕΚ 2014-2020
ΕΣΠΑ 2014-2020
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
clarin:el

Γράφημα t-sne

Ομαδοποίηση των λέξεων με κριτήριο αν οι διανυσματικές τους απεικονίσεις στον γεωμετρικό χώρο είναι κοντά

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στόχοι γραφήματος t-sne

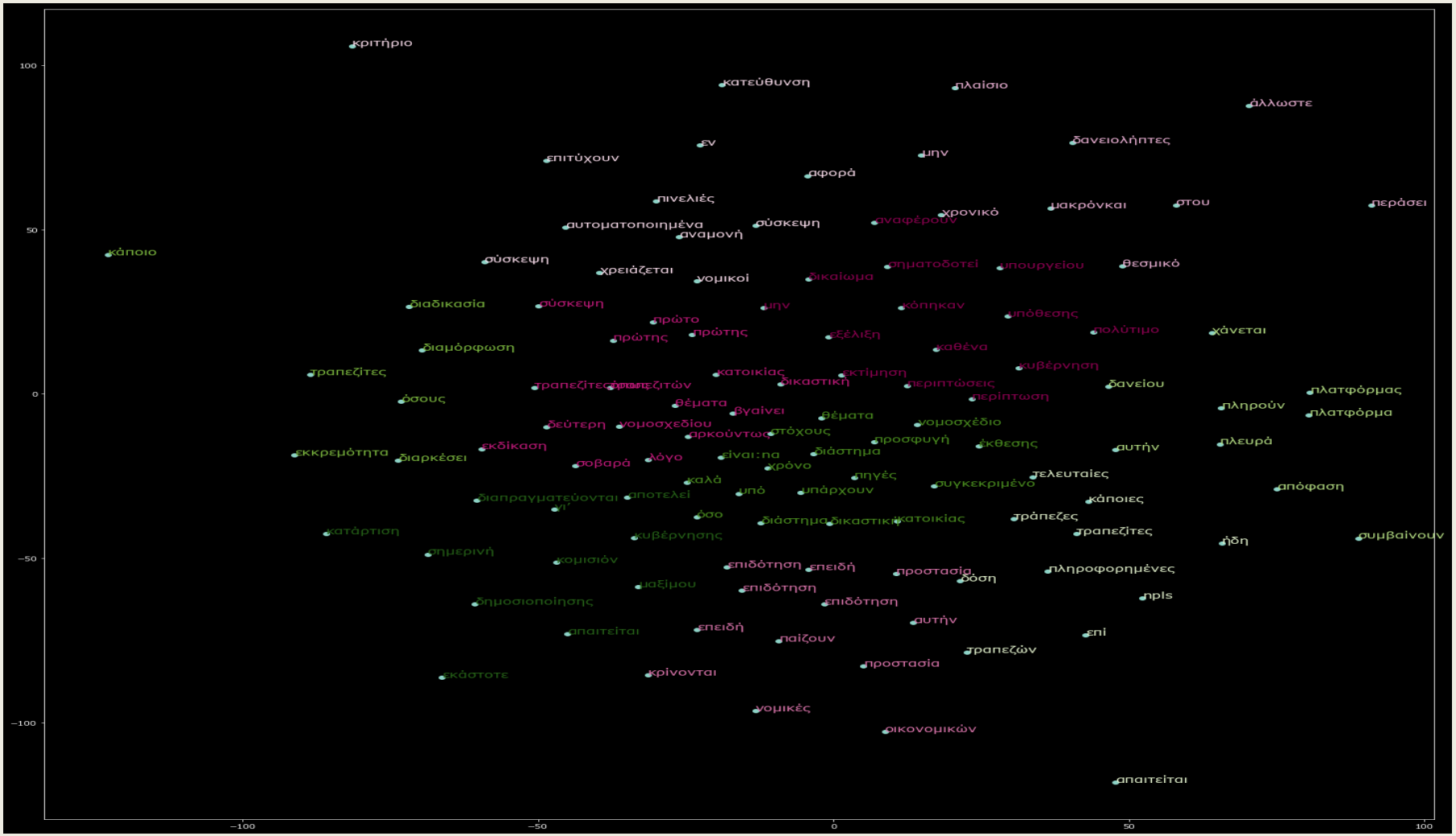
1. Να ενταχθούν στην ίδια ομάδα λέξεις που έχουν τα ίδια χαρακτηριστικά
2. Να αναγνωρισθούν οι σχέσεις μεταξύ των ομάδων των λέξεων
3. Συνόψιση και οπτική αναπαράσταση σημασιολογικών σχέσεων

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

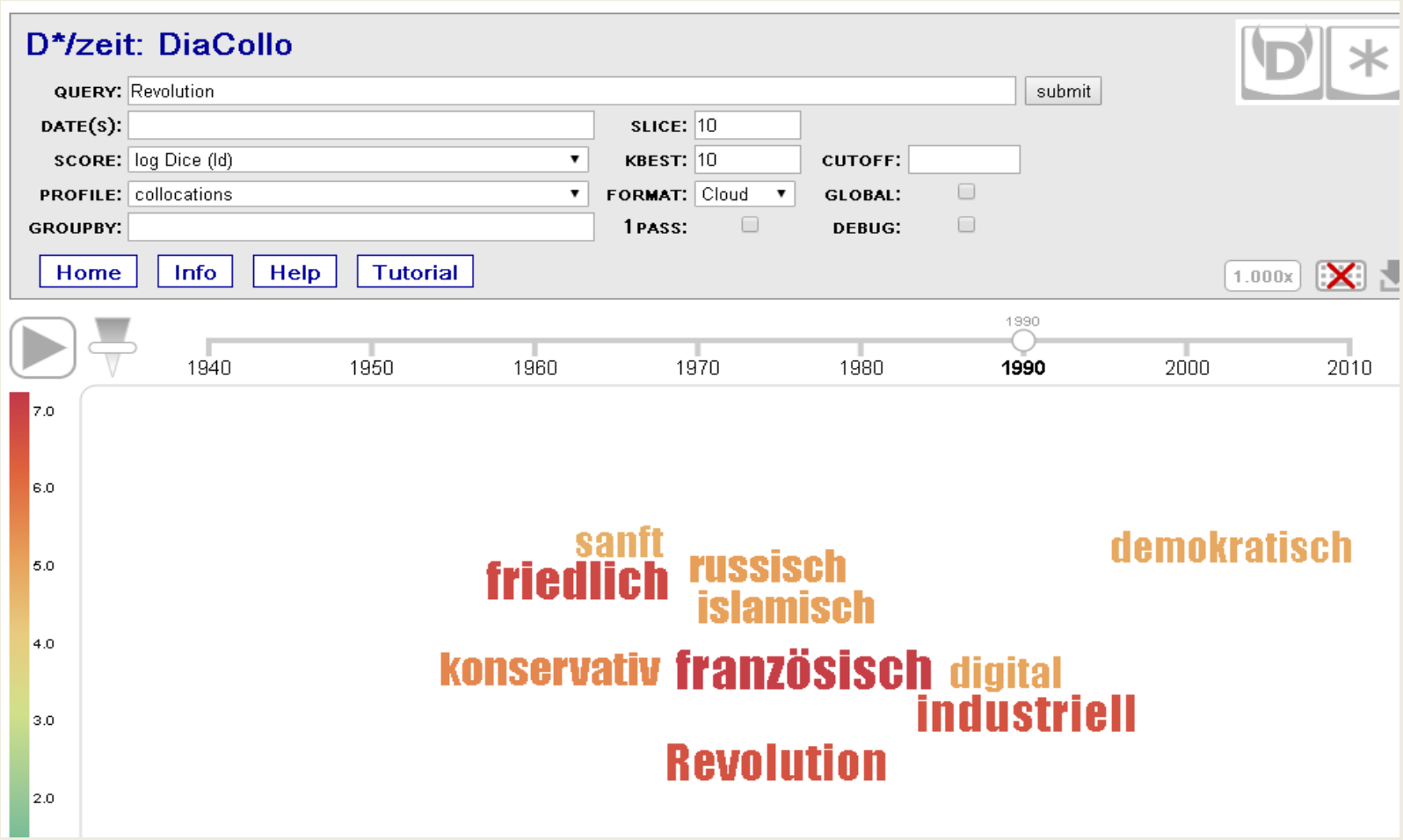
Χαρακτηριστικά γραφήματος t-sne

1. Ο χρήστης μπορεί να επιλέξει μεταξύ τριών μοντέλων (word2vec, fasttext, Glove) για την εκπαίδευση
2. Ο χρήστης μπορεί να αλλάξει τον αριθμό των ομάδων μεταξύ των οποίων θα διαμοιραστούν οι λέξεις του κειμένου (από 2 έως όσες είναι οι λειτουργικές λέξεις του κειμένου
3. Χρήση λίστας stop words του ΣΕΚ

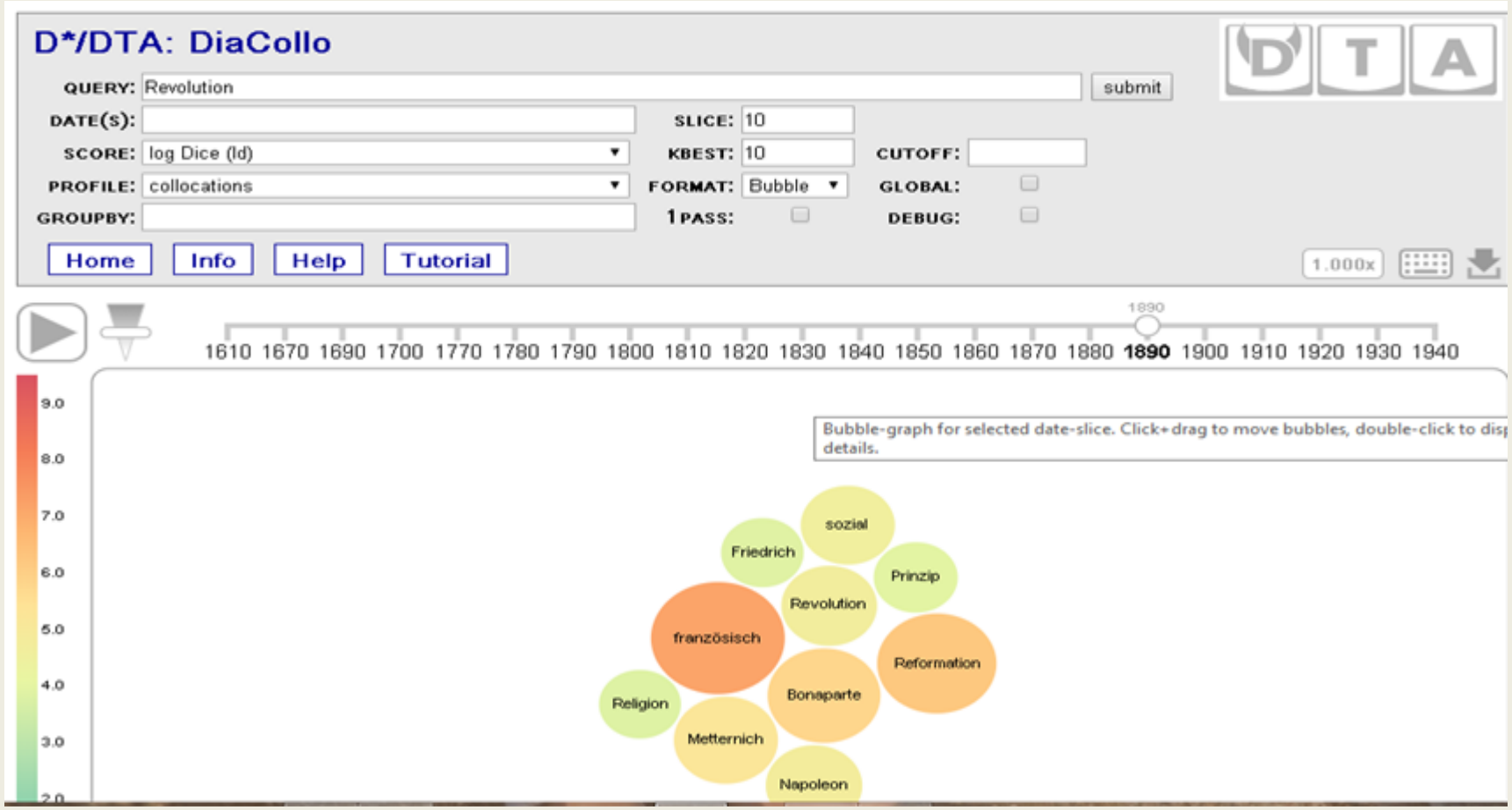
Γράφημα t-sne



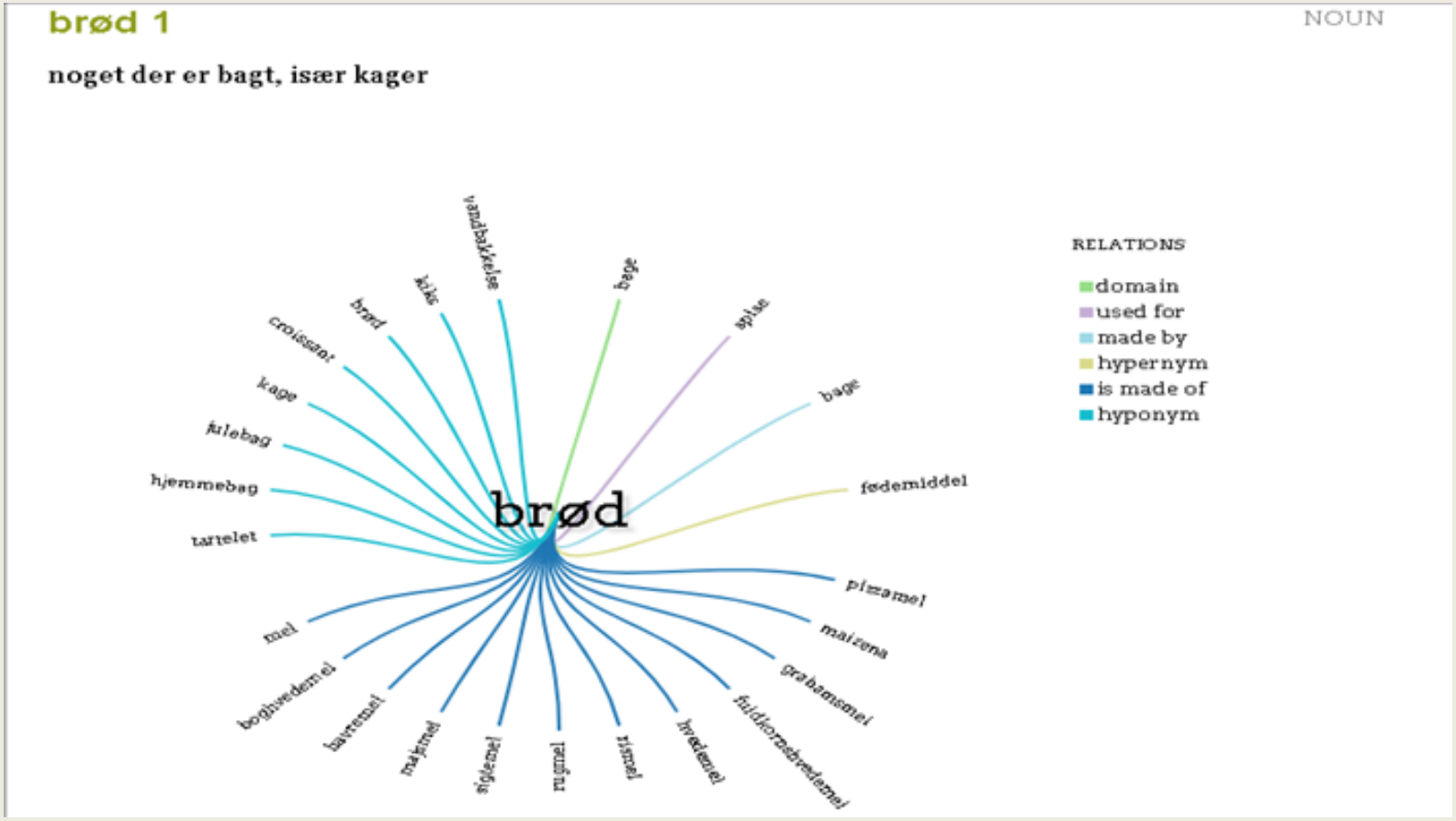
Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες I



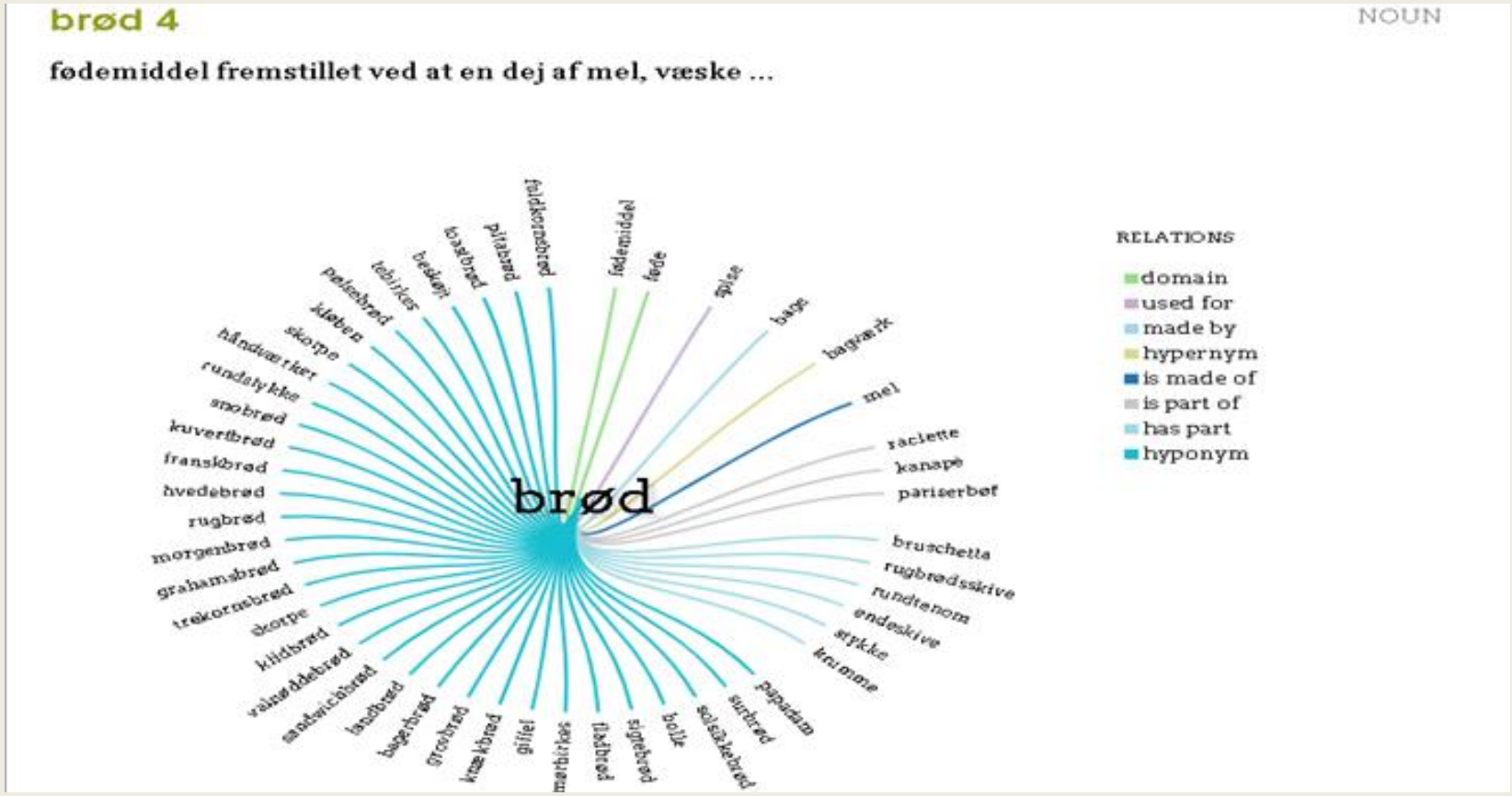
Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες II



Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες III

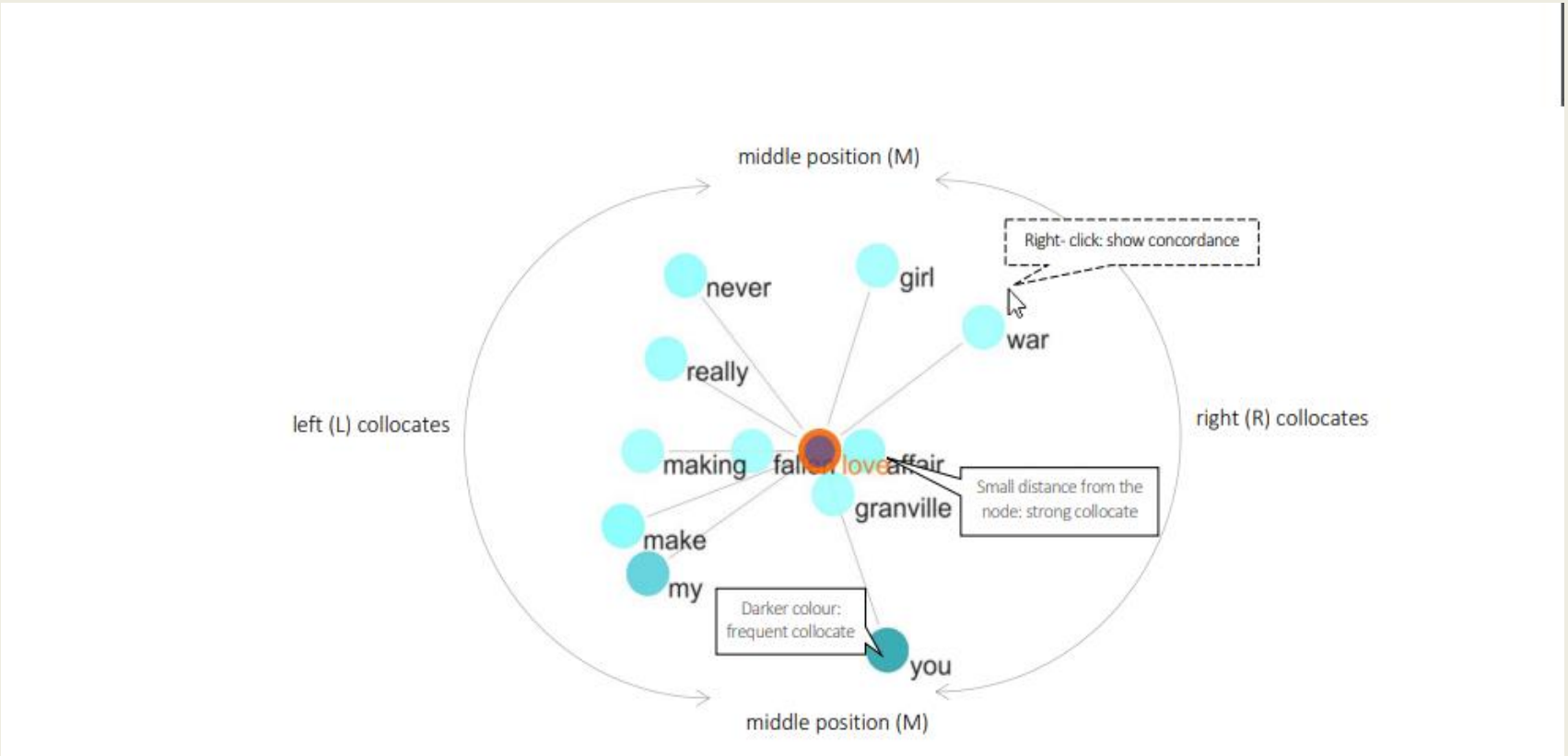


Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες IV

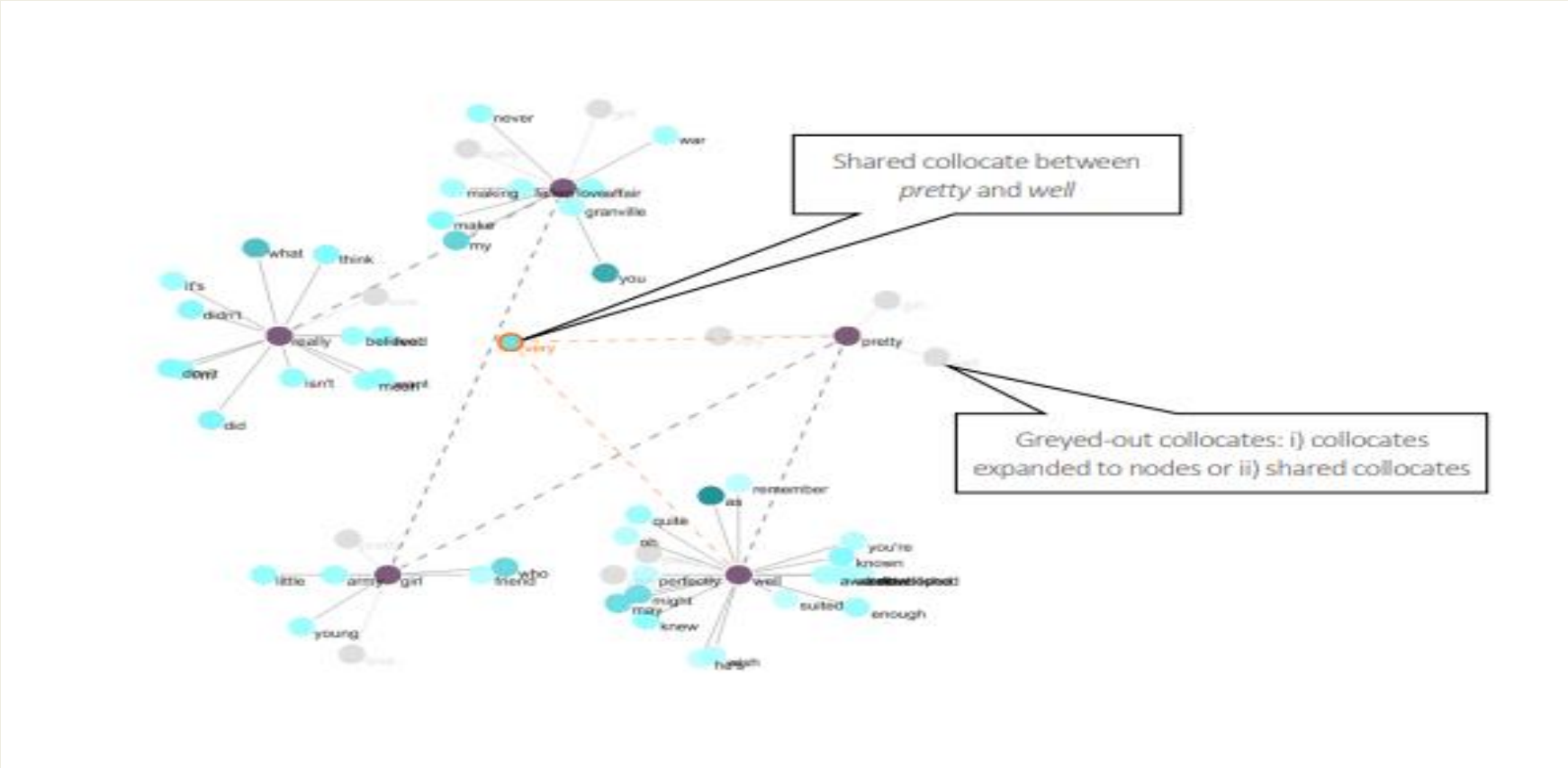


απο λογίς
ΕΠΑνΕΚ 2014-2020
ΕΣΠΑ 2014-2020
clarin:el

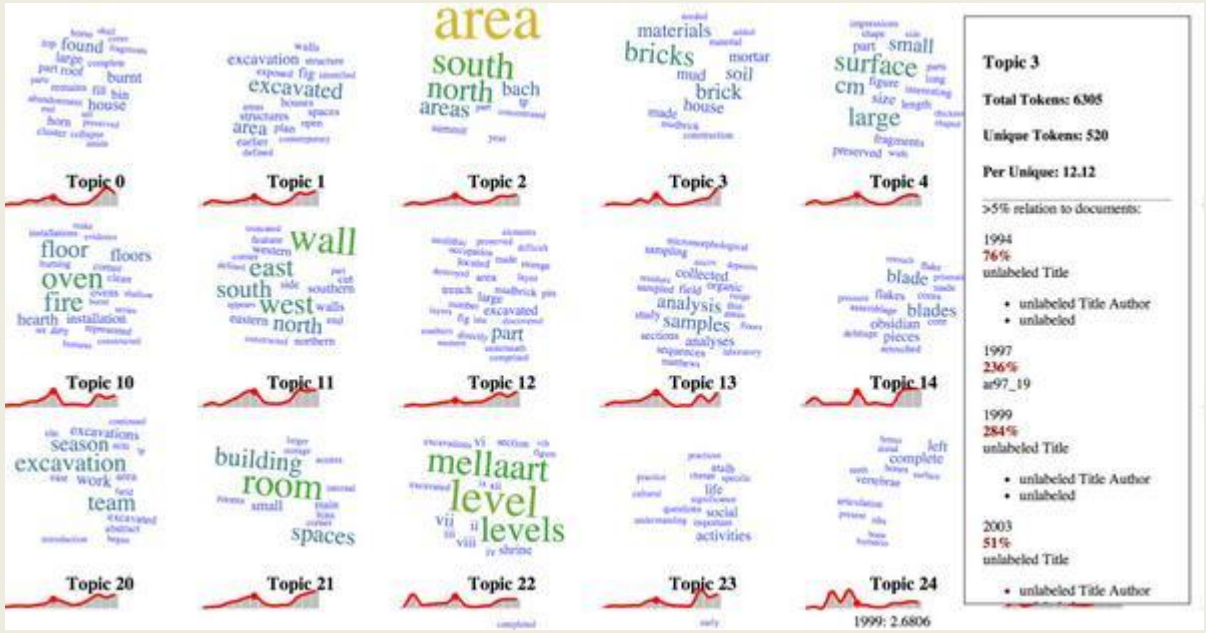
Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες V



Εργαλεία οπτικοποίησης δεδομένων σε άλλες γλώσσες V

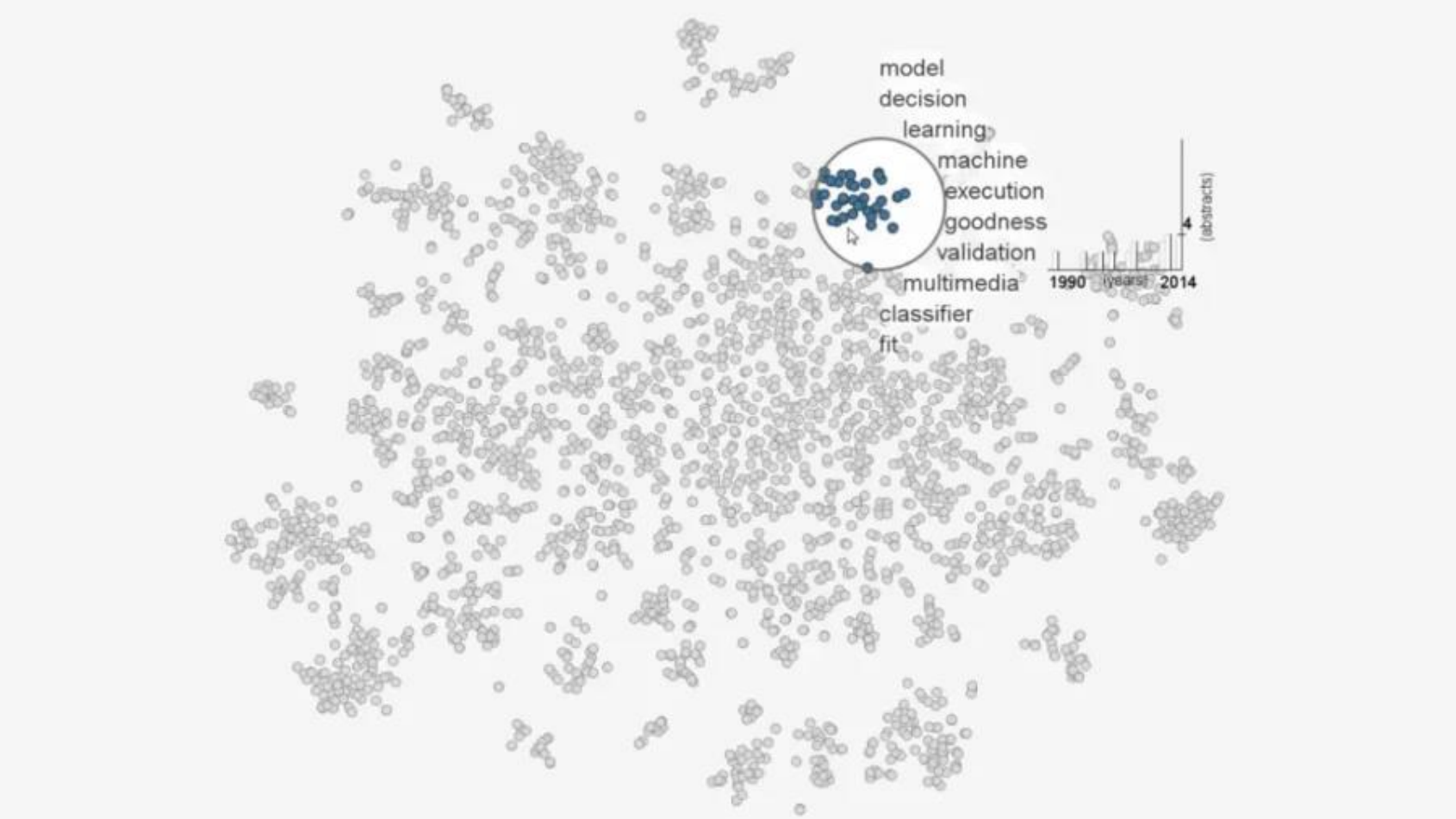


Γλωσσικό εργαλείο (topic visualization) που θα ολοκληρωθεί προσεχώς



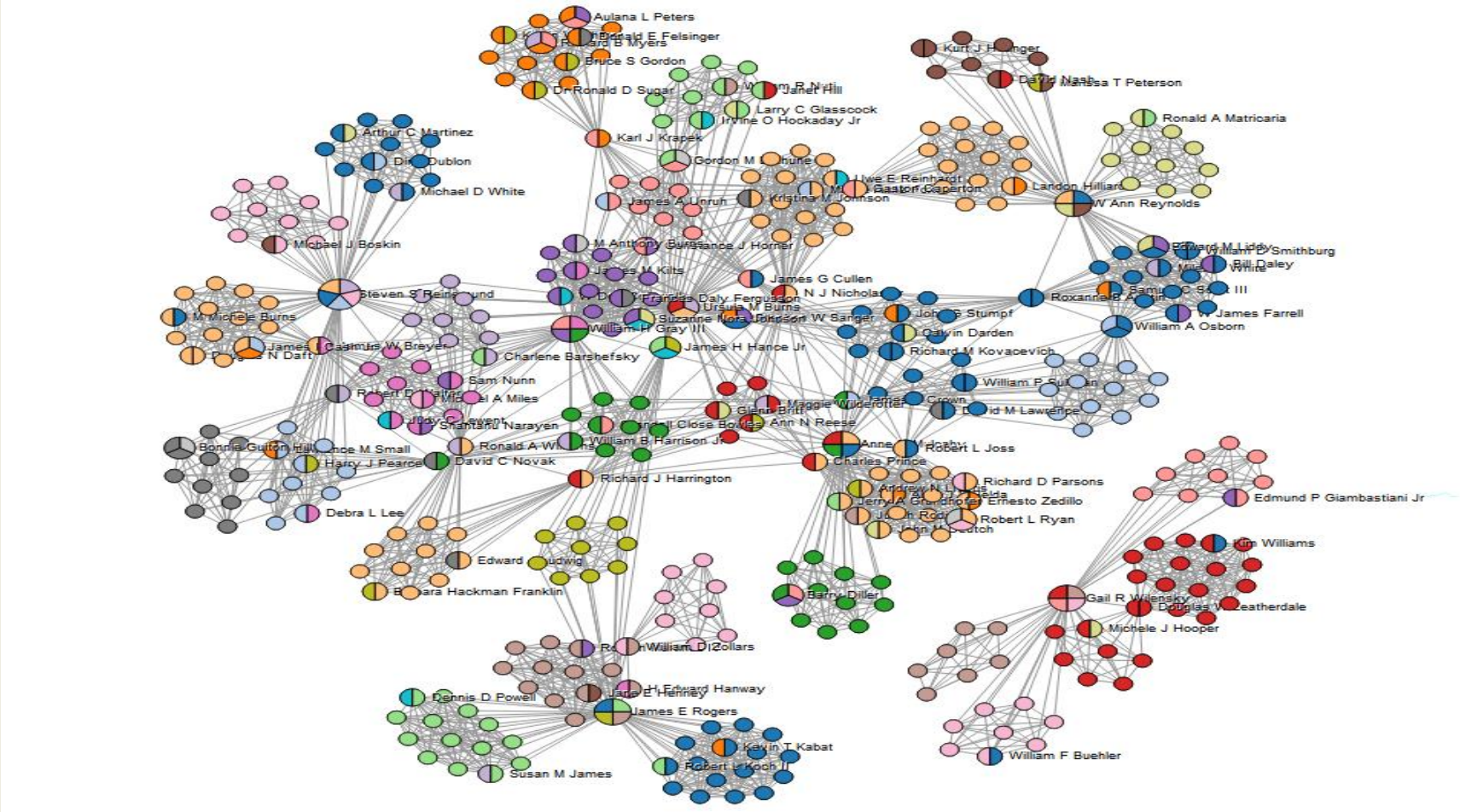
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Γλωσσικό εργαλείο (DocuCompass) που θα ολοκληρωθεί προσεχώς



απο λωνίς
ΕΠΑνΕΚ 2014-2020
ΕΣΠΑ 2014-2020
clarin:el

Γλωσσικό εργαλείο (Graph) που θα ολοκληρωθεί προσεχώς



απολλωνίς
ΕΠΑνΕΚ 2014-2020
ΕΣΠΑ 2014-2020
clarin:el

ΕΥΧΑΡΙΣΤΩ!

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης