Ερμηνευτική ανάλυση λόγου στην ψηφιακή εποχή

Τιτίκα Δημητρούλια
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Manuel Castells

- [Το διαδίκτυο είναι]το φυσικό πεδίο δραστηριοποίησης των σύγχρονων, δυτικών τουλάχιστον, κοινωνιών
- «ο ιστός των ζωών μας» [2002 (2005: 29)]

Ψηφιακότητα και έρευνα: δεδομένα

- Τεράστιος όγκος δεδομένων σε προσβάσιμη μορφή που αφορούν όλες τις επιστήμες (παλαιότερα και νεότερα, ψηφιοποιημένα και ψηφιακά γηγενή)
- Σε διάφορες μορφές: κείμενο, σύνολα δεδομένων (datasets), ήχος, εικόνα, βίντεο, κώδικας κ.λπ.
- Μαζικά ψηφιακά γηγενή δεδομένα προερχόμενα από εφαρμογές του ψηφιακού κόσμου (ιστοσελίδες, κοινωνικά μέσα κ.λπ.)

Ψηφιακότητα και έρευνα: εργαλεία

- Πολλά και διαφορετικά εργαλεία ανάλυσης των δεδομένων σε διαφορετικά επίπεδα
- Διαβαθμισμένης δυσκολίας στη χρήση τους
- Διαφορετικής φιλοσοφίας και στόχευσης

Όλα τα εργαλεία δεν είναι κατάλληλα για κάθε έρευνα

Τα δεδομένα είναι παντού

Αυτό δεν σημαίνει ότι:

- Δεδομένα = γνώση
- Δεδομένα = χρήσιμα δεδομένα
- Όλα τα δεδομένα είναι χρήσιμα για όλες τις επιστήμες
- Σε μια επιστήμη όλα τα δεδομένα μπορούν να αξιοποιηθούν ερευνητικά για όλους τους σκοπούς

Πετρέλαιο ίσον διύλιση = γλωσσική τεχνολογία

Πίσω από τα δεδομένα

- Βάσεις δεδομένων: επίπεδα διαμεσολάβησης σε σχέση με τον χρήστη = οι τεχνολογικές δυνατότητες και οι επιλογές στη δημιουργία τους καθορίζουν την πρόσβασή μας στα δεδομένα
- η πρόσβαση στα δεδομένα και στη διαχείρισή τους ορίζεται από την μορφή της υπηρεσίας (ποικίλοι περιορισμοί)

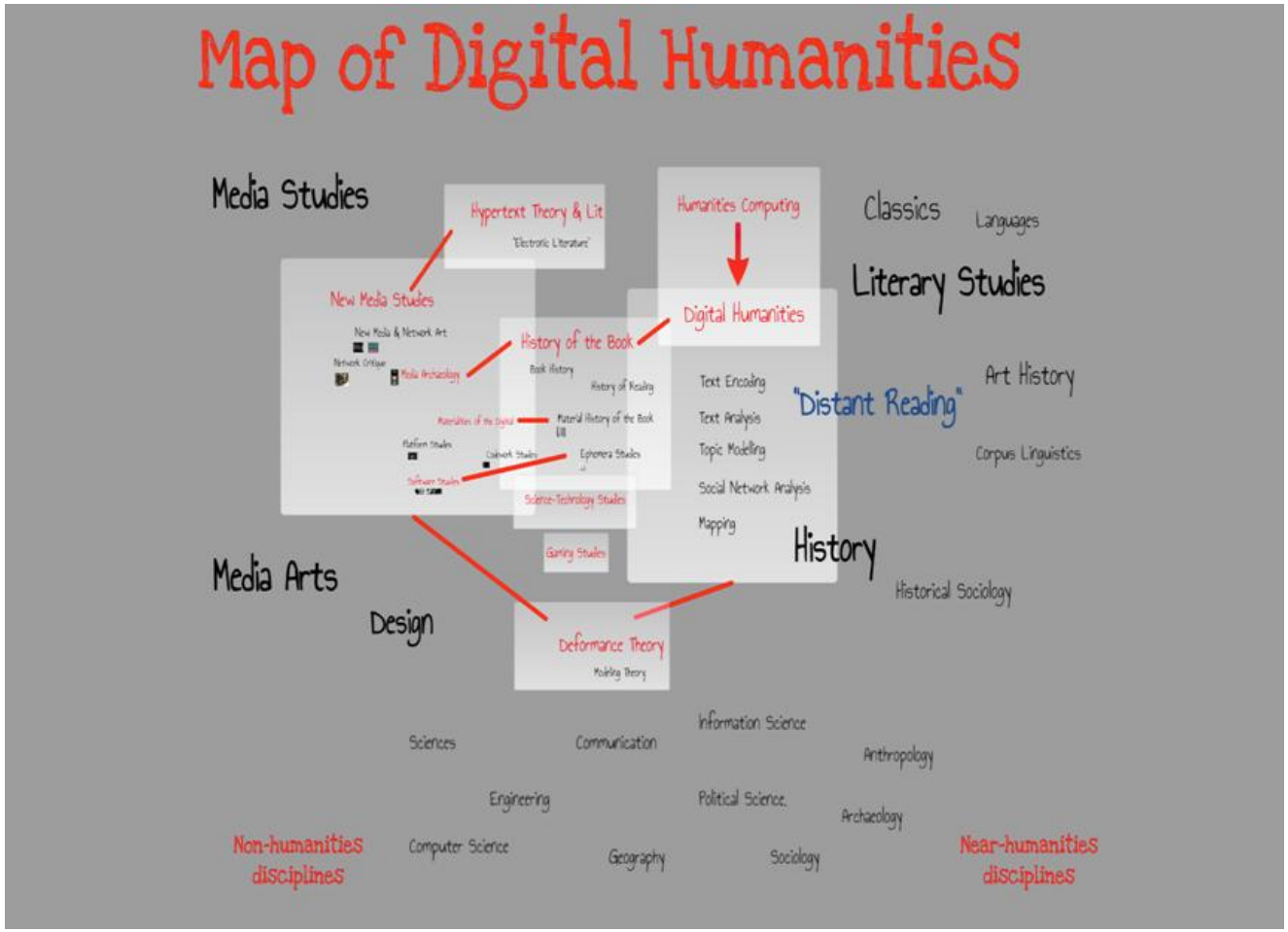
Από τα δεδομένα σημασία και στη γνώση

- **Συλλέγονται** για έναν συγκεκριμένο σκοπό και με συγκεκριμένους τρόπους αναλόγως με τη φύση και τη χρήση τους
- **Αποθηκεύονται** για τους ίδιους λόγους σε συγκεκριμένη μορφή
- Γίνονται αντικείμενο διαφορετικής **επεξεργασίας** αναλόγως με την επιστήμη, το θεωρητικό και μεθοδολογικό μοντέλο που επιλέγεται και το ερευνητικό ερώτημα καθαυτό

Ψηφιακές ανθρωπιστικές επιστήμες

- Τα ζητήματα αυτά τίθενται στις Ψηφιακές Ανθρωπιστικές Επιστήμες
- Που στον γαλλόφωνο χώρο περιλαμβάνουν τις Ανθρωπιστικές, τις Κοινωνικές Επιστήμες και τις Τέχνες
- **Μία μόνο όψη τους είναι η Ηλεκτρονικά υποστηριζόμενη ανάλυση κειμένου** (Computer-Assisted Text Analysis, CATA)
- Αλλά σχετίζονται στο σύνολό τους άρρηκτα με τη γλωσσική τεχνολογία

Αγγλοσαξονική παράδοση: διαφορετικός ορισμός (Alan Liu)



Jerome McGann & Geoffrey Rockwell-Stefàn Sinclair

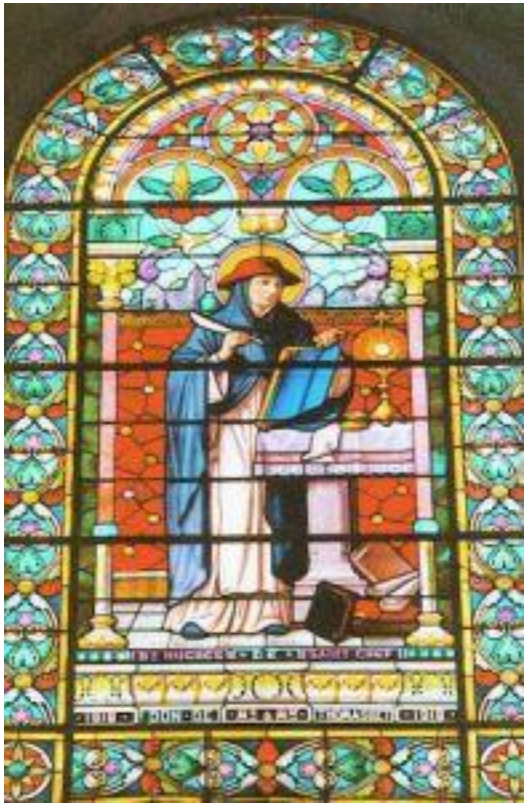
- Ο McGann παραφράζει τον Τόλκιν και βάζει τη λέξη «κείμενο» στη θέση της λέξης «δαχτυλίδι»: το κείμενο είναι «η μία και μόνη λέξη που κυβερνά τους πάντες [ή εδώ τα πάντα] και στα σκοτεινά τους [τα] ρίχνει στα δεσμά» (one word to rule them all […] and in the darkness bind them, 2004:198).
- Οι Rockwell & Sinclair κάνουν λόγο για ubiquitous text, πανταχού παρόν κείμενο

Matthew Kirschenbaum

- πρώτον, μετά τα αριθμητικά δεδομένα, το κείμενο είναι ο πιο πρόσφορος τύπος δεδομένων για επεξεργασία από τους υπολογιστές
- δεύτερον, οι υπολογιστές και η ανάλυση κειμένου συνδέονται στενά εδώ και πολλά χρόνια

Ανάλυση λόγου – ανάλυση κειμένου στον Μεσαίωνα

Ο πρώτος συμφραστικός πίνακας (ασχέτως τεχνολογίας) ανάγεται στον 13ο αιώνα (Huges de Saint-Chef)



Georges Artsrouni, cerveau mécanique, 1933 – μηχανή αποθήκευσης και ανάκτησης-εκτύπωσης αποθηκευμένης πληροφορίας

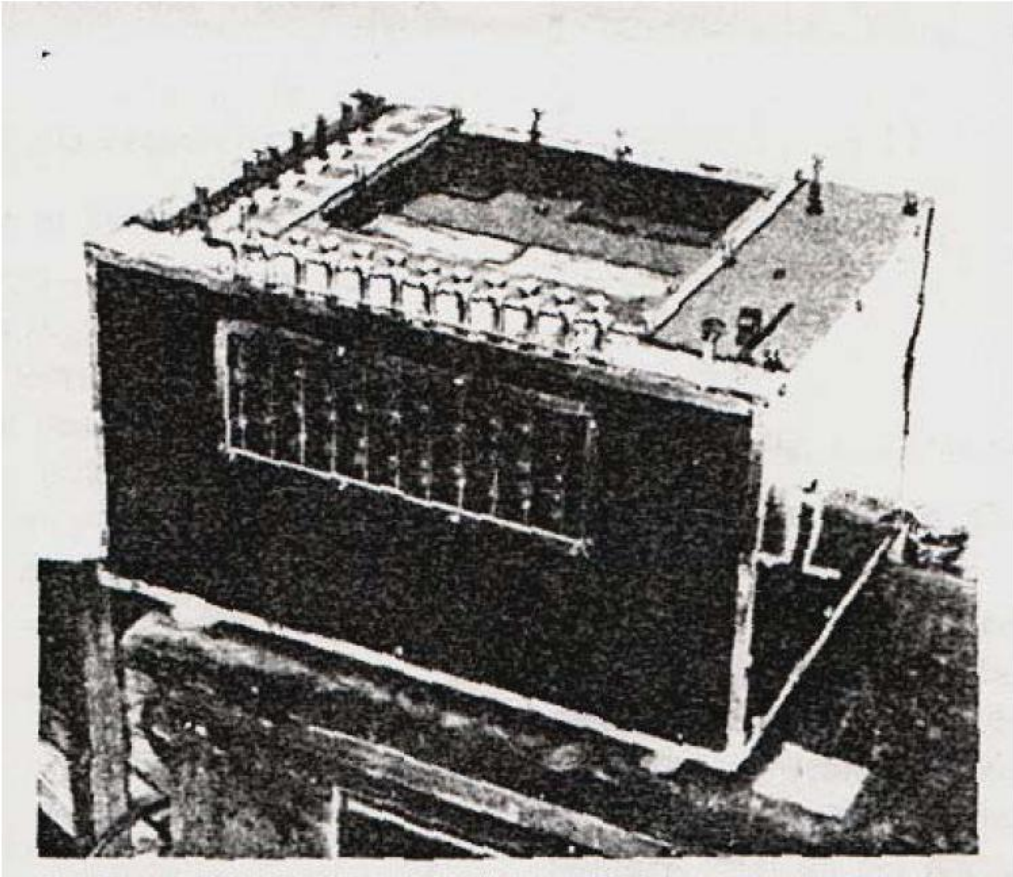


Fig. 1. Artsrouni's machine (*from* Corbé 1960)

Petr Petrovič Trojanskij, 1933, translating machine

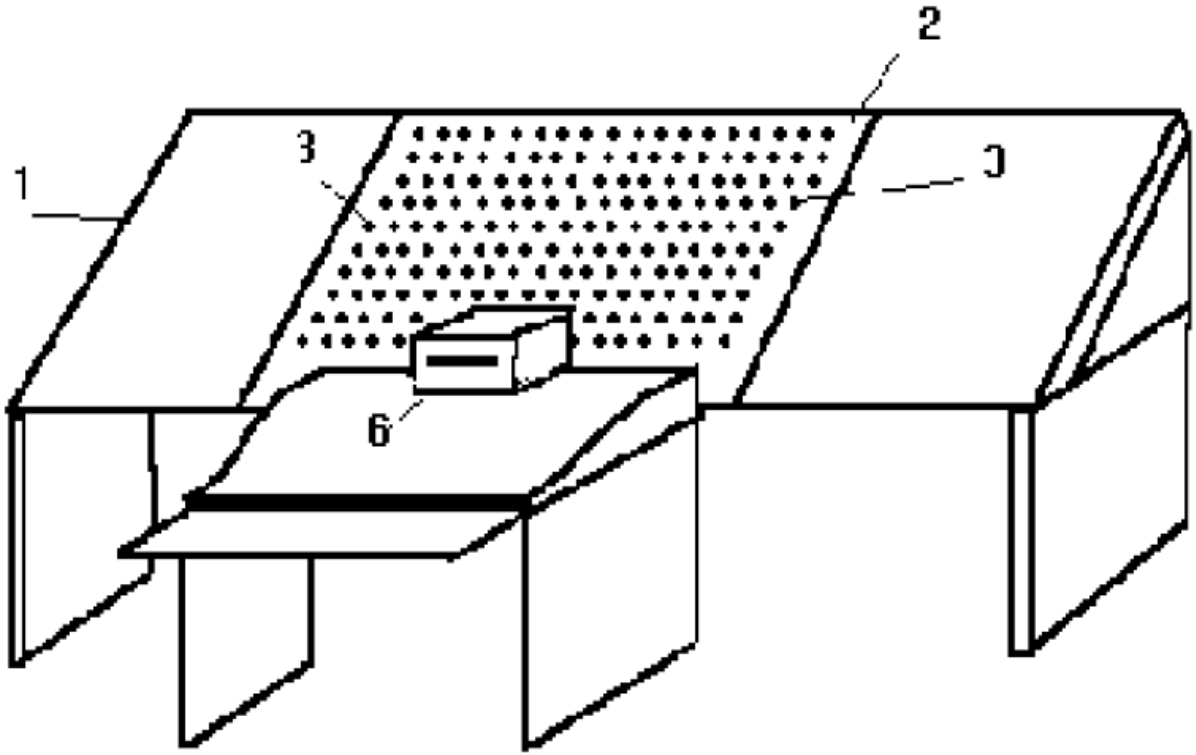


Fig. 3. Trojanskij's translating machine (*from* Bel'skaja et al. 1959)

Index Thomisticus
Πατήρ Busa and IBM 1949-1953

Text(ual) analysis vs Text Mining?

- Άλλοι ερευνητές διαχωρίζουν τις δύο μεθοδολογίες
- Άλλοι όχι

Μια άλλη ταξινόμηση: Wiedemann (2013 και 2015)

1. Εφαρμογές λογισμικού ασχέτως ερευνητικής μεθοδολογίας, που παρέχουν εργαλεία για χειρωνακτική επισημείωση (Computer-Assisted Qualitative Data Analysis - CAQDAS), η οποία λαμβάνει υπόψη τα γλωσσικά συμφραζόμενα και την περίσταση επικοινωνίας

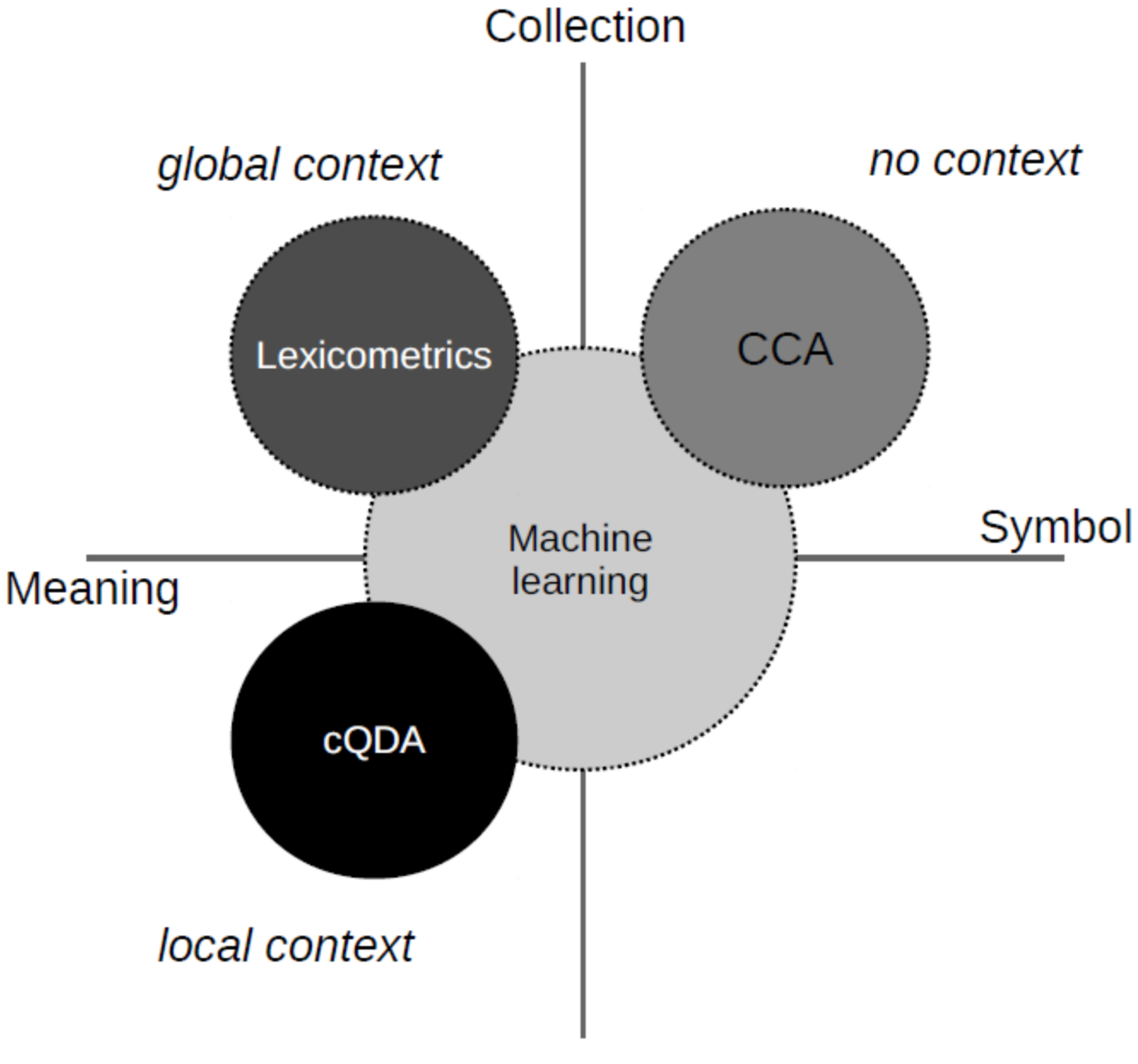
   Ιδιαιτέρως δημοφιλής στις κοινωνικές επιστήμες

2. Ηλεκτρονική ανάλυση περιεχομένου (Computational Content Analysis - CCA), με γνώμονα μια υπόθεση, στην οποία εξετάζονται οι εμφανίσεις όρων σε αυτόματα επισημειωμένα κείμενα χωρίς να λαμβάνονται καθόλου υπόψη τα συμφραζόμενα.

3. Λεξικομετρικές μέθοδοι / μέθοδοι της Γλωσσολογίας Σωμάτων Κειμένων με γνώμονα τα δεδομένα, οι οποίες επιτρέπουν την επαγωγική εξερεύνηση γλωσσικών σχημάτων-μοτίβων, με μέτρηση εμφανών συμφραζομένων.

4. Μέθοδοι εξόρυξης δεδομένων από κείμενα, που επιχειρούν την εξόρυξη «σημασίας» με την εφαρμογή πολύπλοκων στατιστικών μοντέλων, τα οποία υπολογίζουν υπόρρητα συμφραζόμενα.

- Μηχανική μάθηση (machine learning)


Collection
global context
no context
Lexicometrics
CCA
Machine learning
Symbol
Meaning
cQDA
local context
Document

Προαπαιτούμενα της ηλεκτρονικής ανάλυσης κειμένου – ανάλυσης λόγου

- **<u>Ηλεκτρονικά κείμενα</u>**, ψηφιακά ή ψηφιοποιημένα: τεράστια έλλειψη ελληνικού περιεχομένου συν πρόβλημα δικαιωμάτων (παγκόσμιο)
- Γνώση των δυνατοτήτων των εργαλείων (μέχρι νεοτέρας)
- Γνώση των εργαλείων καθαυτά (όλο και πιο εύκολα στη χρήση), εμπορικά και ανοιχτού λογισμικού

Το αντικείμενο του σημερινού σεμιναρίου

- Μία προσέγγιση ανάλυσης κειμένου, λεξικομετρική αλλά όχι μόνο
- Ανάλυση σε πολλά, διαφορετικά επίπεδα, αναλόγως με τις δεξιότητες του χρήστη
- Μπορεί να δώσει άμεσα αποτελέσματα
- Μπορεί να ενσωματώσει δεδομένα επεξεργασμένα με άλλα εργαλεία
- Είναι συλλογική και συνεργατική

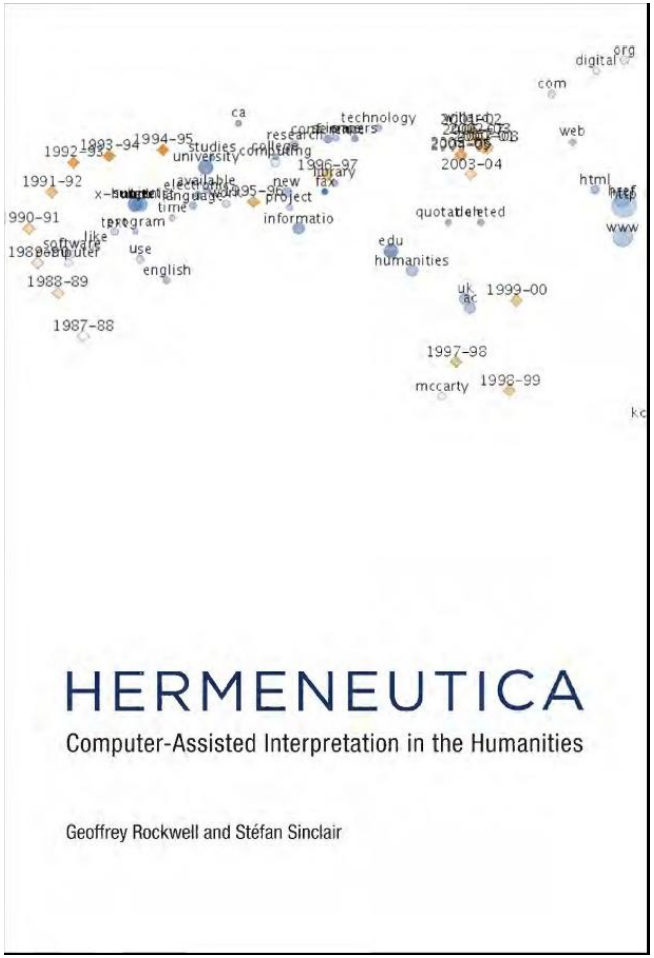
Σημασία ερευνητικού ερωτήματος και μεθόδου

- Ιδιαιτερότητα των ανθρωπιστικών και κοινωνικών επιστημών
- Μια άλλη ματιά στον συνδυασμό ποσοτικού και ποιοτικού στοιχείου: ηλεκτρονική ανάλυση κειμένου
- Ποια ανάλυση;
- Βασισμένη σε θεωρητικές ερμηνευτικές προκείμενες
- Με διαρκή έλεγχο των δεδομένων και της επεξεργασίας τους

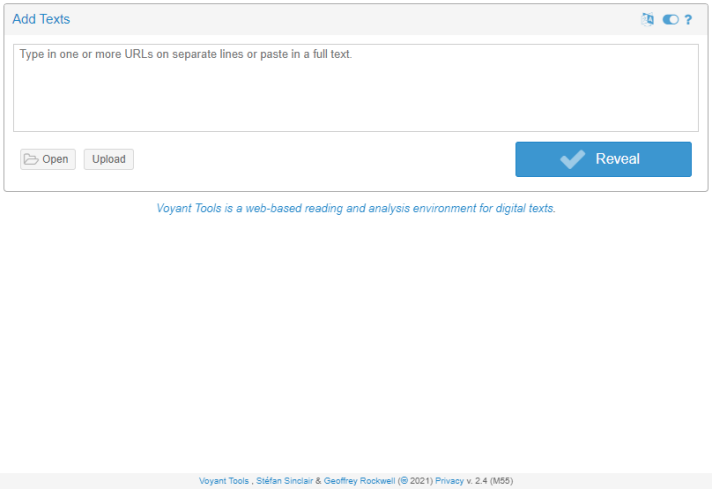
Agile Hermeneutics: μια συνεργατική πρακτική πέραν της ΓΣΚ

- Η έννοια της συνεργασίας: χρησιμοποιώ μια εφαρμογή φτιαγμένη από άλλον ή συνδημιουργώ εφαρμογή = η κοινή δουλειά δεν αποκλείει την ατομική
- Περιορισμοί: ο διαχωρισμός της σύλληψης, της εφαρμογής και της ερμηνείας = μόνη λύση ο **διάλογος και η συνεργασία μεταξύ πολύ διαφορετικών πεδίων**
- Προτεραιότητα: η **ερμηνευτική**

«Μια ιστορία για τις μεθόδους ερμηνείας» hermeneuti.ca



Voyant Tools



Το μείζον ζήτημα, διαχρονικά

… κατά τη γνώμη μου η ανάλυση είναι μια διεργασία που προκύπτει από τη σύγκλιση προηγούμενων διεργασιών. Συνίσταται στη σύνθεση και ανασύνθεση των ιδεών μας, έτσι ώστε να τις συγκρίνουμε και να ανακαλύψουμε, με αυτόν τον τρόπο, τις σχέσεις που έχουν μεταξύ τους, όπως και τις νέες ιδέες που αυτές μπορούν να παραγάγουν…

Geoffrey Rockwell, “What is Text Analysis, Really?”, Literary and Linguistic Computing, Vol. 18, No. 2, 2003, p. 209-219

ÉTIENNE BONNOT DE Condillac Essai sur l’Origine des connaissances humaines, CH. HOUEL, IMPRIMEUR, PARIS, 1798

Ποιο κείμενο;
Ταυτόν τε και θάτερον;

ΑΤΘΙΔΕΣ ΑΥΡΑΙ.

ΟΙ ΓΑΜΟΙ.

Ἀπ' ἄκρη σ' ἄκρη διαλαλεῖ
ὁ Γέρο-κοῦκκος πάλι,
κ' εἰς ἕνα γάμο προσκαλεῖ
κάθε πουλὶ ποῦ ψάλλει.

Κ' οἱ καλεσμένοι φτερωτοὶ
στὴν ἐξοχὴ τραβοῦνε,
γιὰ νὰ χαροῦνε τὴν γιορτή,
τοὺς γάμους νὰ χαροῦνε.

Φοροῦν τὰ δένδρα γιορτερὰ
μοσχάτα τὰ λουλούδια,
καὶ τὰ φυτὰ τὰ δροσερὰ
διαμάντινα πρεπούδια.

B

Οἱ Γάμοι

Συγγραφέας: Γεώργιος Βιζυηνός

Ἀπ' ἄκρη σ' ἄκρη διαλαλεῖ
ὁ Γέρο-κοῦκκος πάλι,
κ' εἰς ἕνα γάμο προσκαλεῖ
κάθε πουλὶ ποῦ ψάλλει.

Κ' οἱ καλεσμένοι φτερωτοὶ
στὴν ἐξοχὴ τραβοῦνε,
γιὰ νὰ χαροῦνε τὴν γιορτή,
τοὺς γάμους νὰ χαροῦνε.

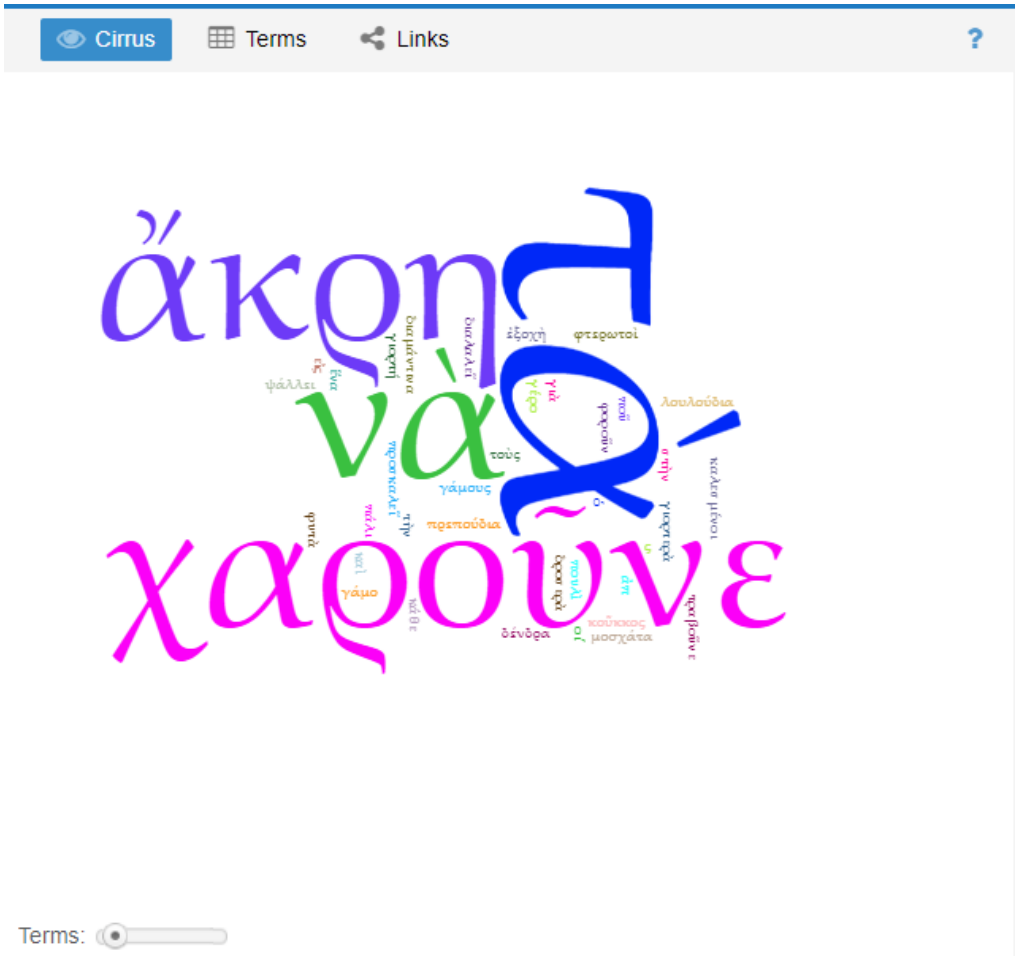
Φοροῦν τὰ δένδρα γιορτερὰ
μοσχάτα τὰ λουλούδια,
καὶ τὰ φυτὰ τὰ δροσερὰ
διαμάντινα πρεπούδια.

Hybrid, monster, bug of words

Cirrus | Terms | Links

	Term	Count	Trend
1	τὰ	4	
2	ἄκρη	2	
3	νὰ	2	
4	χαροῦνε	2	
5	ἀπ	1	
6	γάμο	1	
7	γάμους	1	
8	γέρο	1	
9	γιὰ	1	
10	γιορτερὰ	1	
11	γιορτή	1	
12	δένδρα	1	
13	διαλαλεῖ	1	
14	διαμάντινα	1	

40


Cirrus
Terms
Links
?
ἄκρη
νὰ
χαροῦνε
ἐξοχή
φτερωτοί
λουλούδια
τοὺς
γάμους
πρεπούδια
γάμο
δένδρα
κοῦκκος
μοσχάτα
Terms:

Ηλεκτρονικά Σώματα Κειμένων

- Corpus / corpus-based /corpus-driven / corpus-assisted analysis, study, approach etc.
- DIY (Do It Yourself)corpora
- Διαθέσιμα ΗΣΚ
- Αποθετήρια γλωσσικών πόρων και εργαλείων
- Αυθεντικά δεδομένα
- Υπέρβαση της υποκειμενικής εμπλοκής
- Διαχείριση μαζικών δεδομένων
- Ποιοτική ανάλυση των ποσοτικών στοιχείων

Πληθώρα εργαλείων αλλά δεν κάνουν όλα την ίδια δουλειά!

digitalresearchtools

log in help

DiRT PLANT SEEDS. GROW IDEAS. DIGITAL RESEARCH TOOLS WIKI

Wiki

Search this workspace

If you are citizen of an European Union member nation, you may not use this service unless you are at least 16 years old.

Finally, you can manage your **Google Docs**, uploads, and email attachments (plus **Dropbox** and **Slack** files) in one convenient place. Claim a free account, and in less than 2 minutes, **Dokkio** (from the makers of PBworks) **can automatically organize your content for you.**

VIEW

Text Analysis Tools

DiRT (Digital Research Tools) has a new home! Please visit Bamboo DiRT to explore this excellent collection of research tools.

Definition:

Text analysis software enables users to determine the frequency with which words or phrases are used, create concordances, view words in context, and otherwise study patterns in texts.

Tools:

- AntConc 3.2.1: concordance program, "can generate KWIC concordance lines and concordance distribution plots...has tools to analyze word clusters (lexical bundles, n-grams, collocates, word frequencies, and keywords" (Free, Windows/Mac OS X/Linux)
- Attensity: text analytic and data extraction framework: "delivers the power of sophisticated data and semantic analytics in an integrated suite of easy-to-use business applications, allowing business leaders, customer support personnel and customers to get relevant and actionable answers fast" (Commercial, cross-platform)
- Basis Technology: natural language processing technology for the analysis of unstructured multilingual text (Commercial, cross-platform)
- CATMA (Computer Aided Textual Markup and Analysis): " an open source software with a focus on textual markup and analysis." (Open source)
- ClearForest: text tagging, extraction and analytics software; "Compare Suite compares texts by keywords, highlights common and unique keywords. Connexor Machine discovers natural language grammatical and semantic information." (Commercial)
- Coding Analysis Toolkit (CAT): a platform-independent, open source, system which consists of a web-based suite of tools custom built to facilitate efficient and effective analysis of text datasets that have been coded using either an internal coding module or the commercial-off-the-shelf package ATLAS.ti; CAT computes interrater reliability and supports adjudication and validty measurement (Free, web-based)
- Cypher: a "software program available which generates the RDF graph and SPARQL/SeRQL query representation of a plain language input, allowing users to speak plain language to update and query semantic databases...With robust definition languages, Cypher's grammar and lexicon can quickly and easily be extended to process highly complex sentences and phrases of any natural language, and can cover any vocabulary" (Free, Windows/Mac OS X/Linux)
- Data for Research (DfR): Mine and analyze JSTOR's collections. Supports fielded searching; provides ngrams, word frequencies, citations, and tag clouds of key terms; offers API for "content selection and retrieval." (Free, web-based)
- DICTION 5.0: computer-aided text analysis for determining the tone of a verbal message: certainty, activity, optimism, realism, and commonalty (Commercial, Windows)
- DiscoverText.com: import data, search and analyze, code documents, generate reports. (Commercial, cloud-based)
- Edition Production & Presentation Technology (EPPT): "an integrated set of XML tools designed to help humanities editors prepare image-based electronic editions...makes image-based encoding, the laborious process of linking descriptive markup to material evidence through XML, a relatively easy and error-proof task" (Free, PC XP/Mac OS X)
- HyperPo: "a user-friendly text exploration and analysis program"; supports word frequencies, KWIC (Keyword in Context), cooccurrence and distribution lists, comparison, etc. (Free, web-based)
- IBM AeroText: "a suite of text mining applications that are used for content analysis...Sample target applications include automatic database generation, document routing, browsing, summarization, enhanced full text search, and targeted document search in addition to link analysis" (Commercial, Windows/Linux/Solaris)
- IBM InfoSphere: a product line from IBM which "provides a unified data warehouse delivering access to structured and unstructured information and operational and transactional data in real time" (Commercial, Windows/Linux/UNIX)
- ICTA: "a web-based system for **Automated Text Analysis and Discovery of Social Networks** from text. It was originally designed to work with email-based and forum-based data. But it can also be used to analyze other types of electronic communication such as blogs and chats." (Free, web-based)
- JGAAP: "Java-based, modular program for textual analysis, text categorization, and authorship attribution" (Free)
- Juxta: "tool for comparing and collating multiple witnesses to a single textual work. The software allows users to set any of the witnesses as the base text, to add or remove witness texts, to switch the base text at will, and to annotate Juxta-revealed comparisons and save the results." (Free, cross-platform)
- Lextek: offers a range of services and software for full-text indexing search and retrieval; automatic classification, routing, and filtering electronic text according to user defined profiles (Commercial, cross-platform)
- LIWC: "Linguistic Inquiry and Word Count (LIWC) is a text analysis software program...LIWC is able to calculate the degree to which people use different categories of words across a wide array of texts." (Commercial, with free and limited web analysis available; Windows/Mac)
- MALLET: "a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text." (open source using the Common Public License, Java-based)
- MAXQDA: a tool for qualitative data analysis, evaluation, and text analysis: "supports all individuals performing qualitative data analysis and helps to systematically evaluate and interpret texts...also a powerful tool for developing theories and testing the theoretical conclusions of the analysis" (Commercial, Windows)
- MONK: "a digital environment designed to help humanities scholars discover and analyze patterns in the texts they study. It supports both micro analyses of the verbal texture of an individual text and macro analyses that let you locate texts in the context of a large document space consisting of hundreds or thousands of other texts" (Free, web-based)
- MonoConc: a "concordance (text searching) program...used in the analysis of English or other texts...also produces wordlists and collocation information" (Commercial, Windows)
- MorphAdorner: "a Java command-line program which acts as a pipeline manager for processes performing morphological adornment of words in a text...Currently MorphAdorner provides methods for adorning text with standard spellings, parts of speech and

Tags: HyperPo, JGAAP, Juxta, MAXQDA, MONK, MonoConc, NORA, NVivo, PhiloLogic, Stanford POS Tagger, TACTWeb, TAMS Analyzer, TAPOR Tools, TagCrowd, Token-X, WMatrix, template, text analysis, text mining, text visualization

Check for plagiarism

To join this workspace, request access.

Already have an account? Log in!

This version of DiRT has been superseded by Bamboo DiRT, developed by Quinn Dombrowski and Project Bamboo. Bamboo DiRT makes several improvements over the old DiRT. No new information will be added here as of 1/9/2012, but this wiki is offered for historical purposes.

Authoring
Blogging
Brainstorming
Citation Management
Collaborative Authoring
Collection Building
Communication Tools
Comparison Tools
Data Collection
Data Mining
Data Visualization
Find Research Materials
Image Editing
Linguistic Tools
Mapping
Mashups
Mobile Devices
Networking
Notetaking
Organize Research Materials
Qualitative Data Analysis
Screencasts
Social Bookmarking
Statistical analysis

Ce que le numéri....html

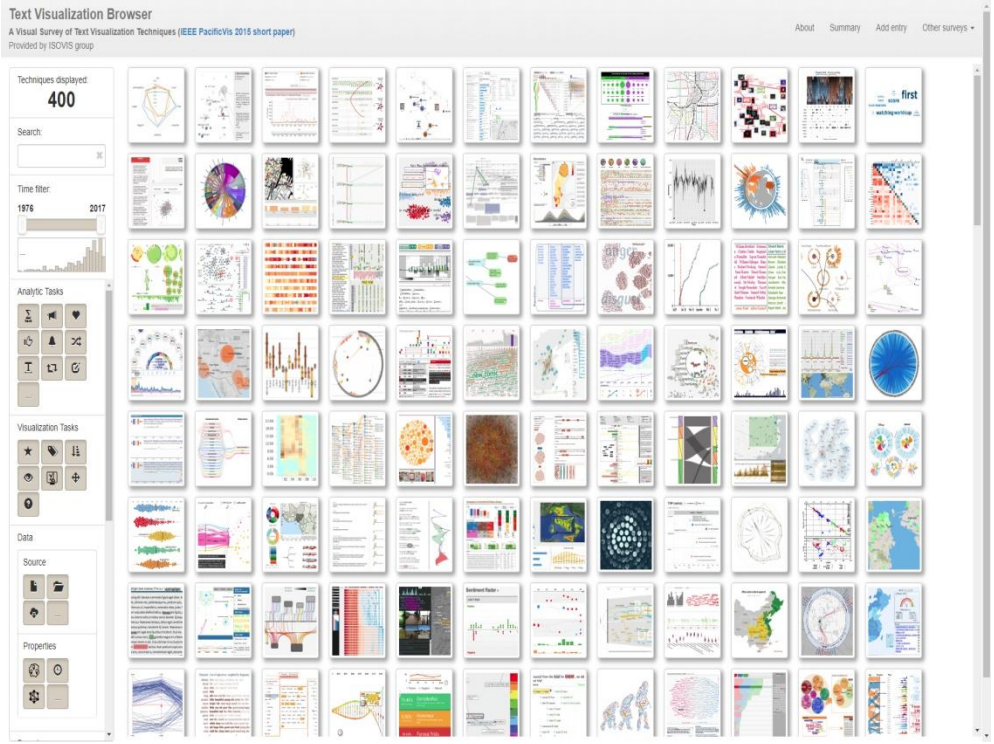
Show all

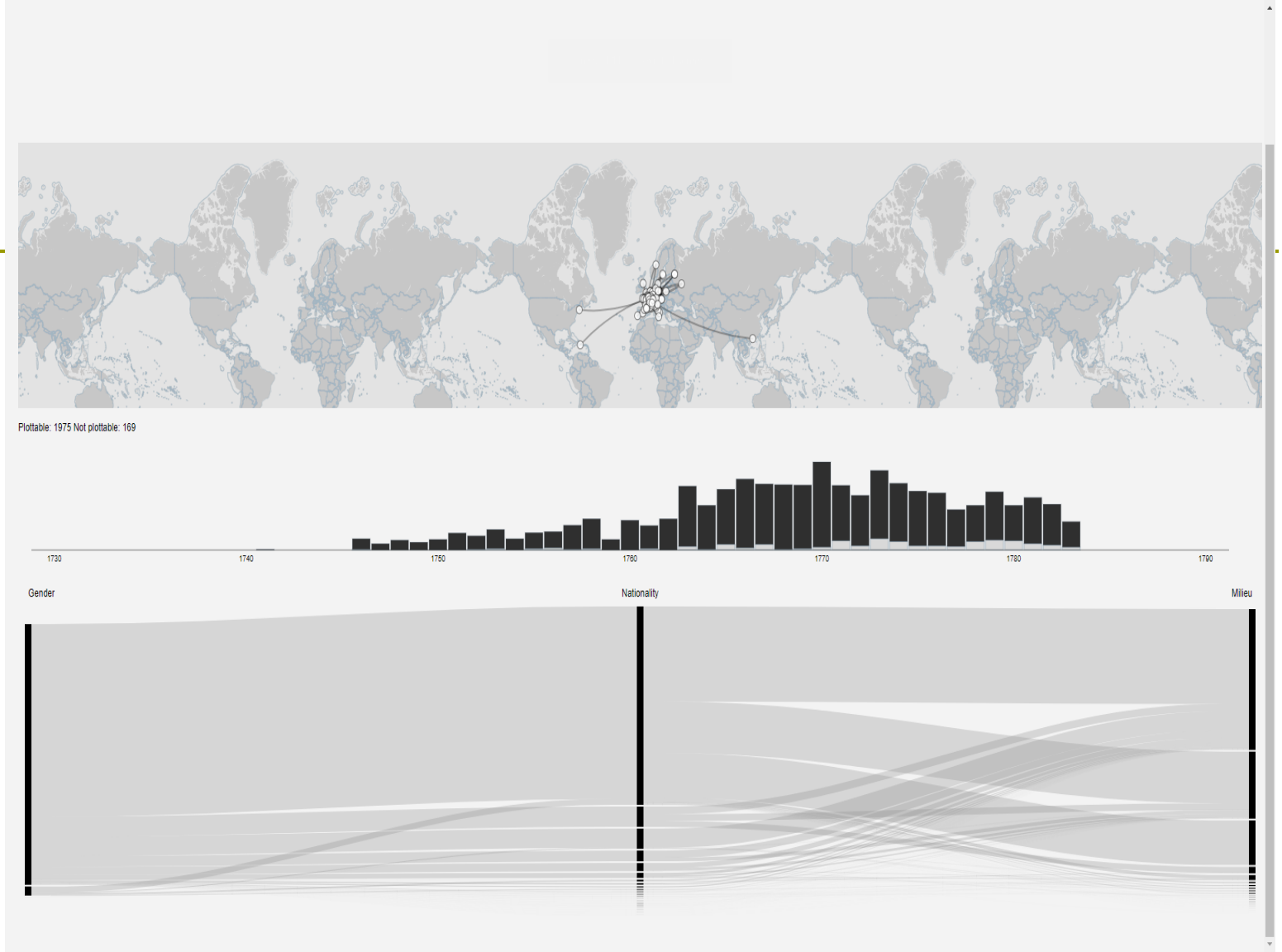
https://digitalresearchtools.pbworks.com/w/page/17801708/Text-Analysis-Tools

Τι μας δίνει ο υπολογιστής

- Τεράστια μνήμη
- Συνδυαστικές δυνατότητες
- Διαχείριση τεράστιου όγκου δεδομένων
- Ποσοτικά στοιχεία (αλλά όχι μόνο)

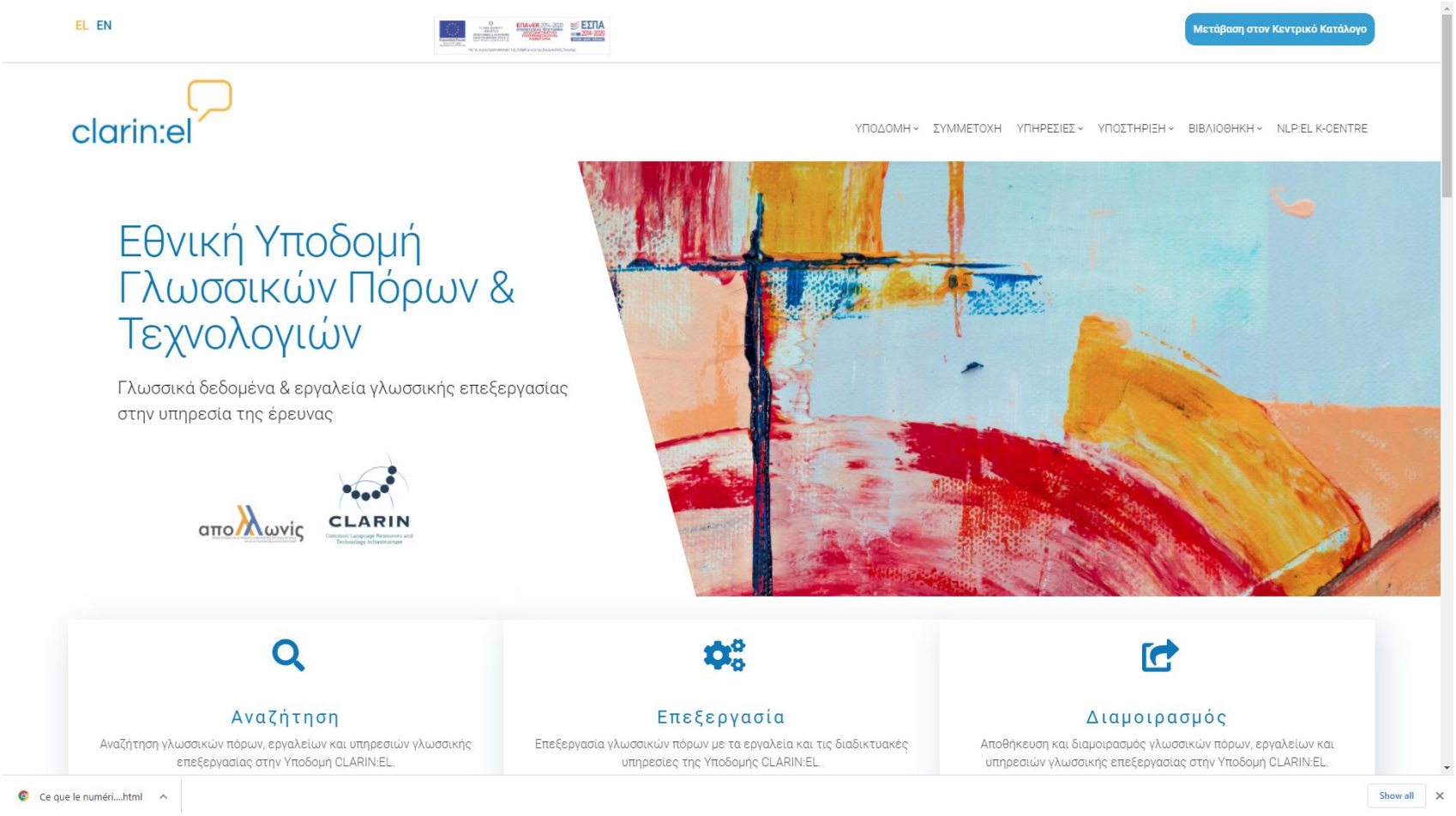
Από το κείμενο στην εικόνα
400 τρόποι οπτικοποίησης



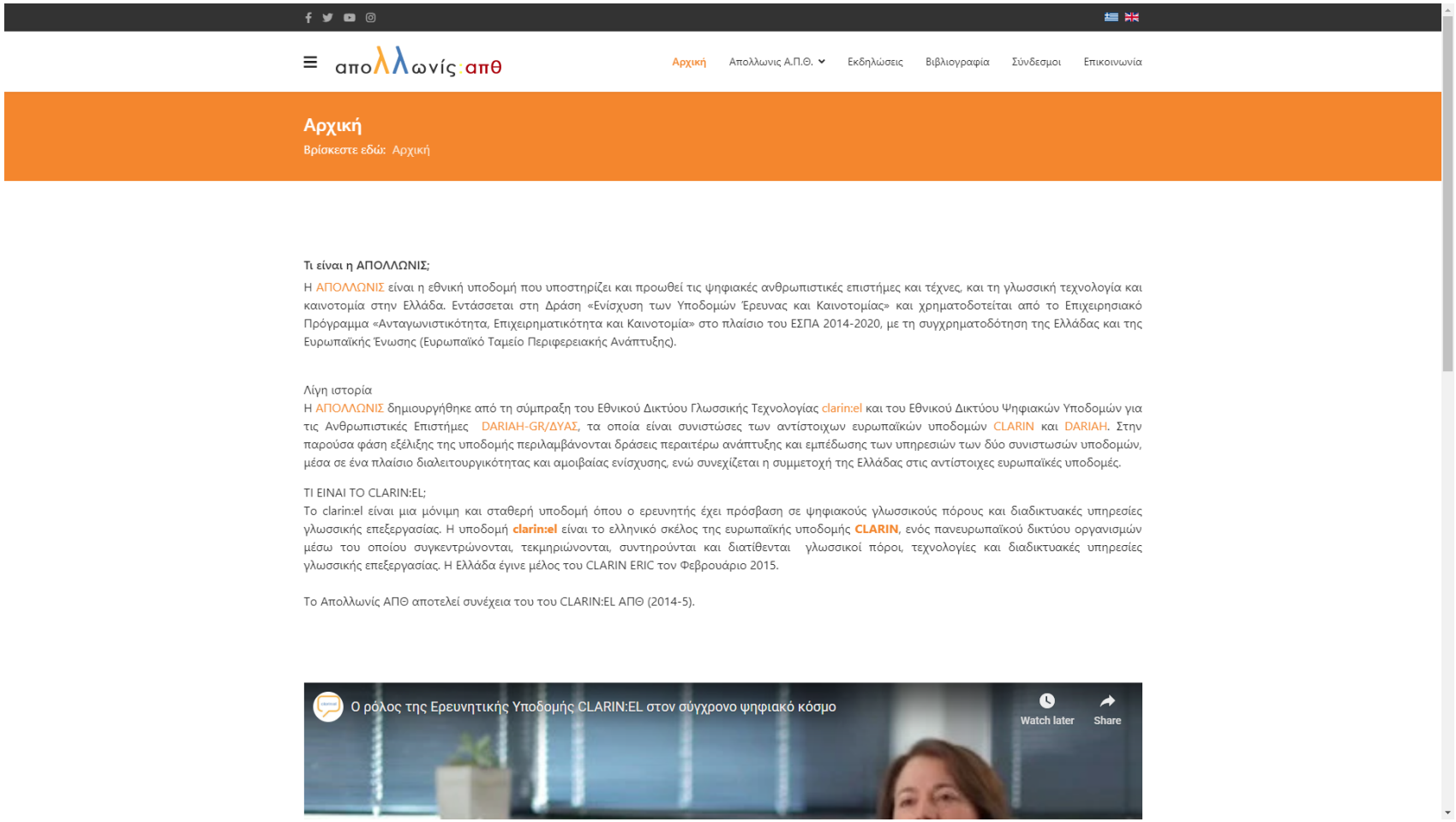


http://ink.designhumanities.org/dalembert/

Clarin


EL EN
Μετάβαση στον Κεντρικό Κατάλογο
clarin:el
ΥΠΟΔΟΜΗ ΣΥΜΜΕΤΟΧΗ ΥΠΗΡΕΣΙΕΣ ΥΠΟΣΤΗΡΙΞΗ ΒΙΒΛΙΟΘΗΚΗ NLP:EL K-CENTRE
Εθνική Υποδομή Γλωσσικών Πόρων & Τεχνολογιών
Γλωσσικά δεδομένα & εργαλεία γλωσσικής επεξεργασίας στην υπηρεσία της έρευνας
CLARIN
Αναζήτηση
Αναζήτηση γλωσσικών πόρων, εργαλείων και υπηρεσιών γλωσσικής επεξεργασίας στην Υποδομή CLARIN:EL.
Επεξεργασία
Επεξεργασία γλωσσικών πόρων με τα εργαλεία και τις διαδικτυακές υπηρεσίες της Υποδομής CLARIN:EL.
Διαμοιρασμός
Αποθήκευση και διαμοιρασμός γλωσσικών πόρων, εργαλείων και υπηρεσιών γλωσσικής επεξεργασίας στην Υποδομή CLARIN:EL.
Ce que le numéri....html
Show all

Απολλωνίς ΑΠΘ



29.3.2021 "Ερμηνευτική ανάλυση λόγου στην ψηφιακή εποχή"

Πληθώρα αλλαγών

Αλλά στο επίκεντρο πάντα ο λόγος και ο μελετητής του

Το μέλλον είναι εδώ!

Βιβλιογραφία

- Busa, Roberto. “The Annals of Humanities Computing: The Index Thomisticus”. *Computers and the Humanities* 14 (1980): 83–90.
- Castells, Manuel, *Ο γαλαξίας του διαδικτύου.* Αθήνα: Καστανιώτης, 2005.
- Drucker, Johanna. “Data Mining and Text Analysis”. In *Intro to Digital Humanities* (online coursebook), 2014, http://dh101.humanities.ucla.edu/?page_id=48
- Kirschenbaum, Matthew. “What is Digital Humanities and What’s It Doing in English Departments”. *ADE Bulletin* 150 (2010): 55-61.
- Rockwell, Geoffrey and Stefàn Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge,MA: MIT Press, 2016.
- Rockwell, Geoffrey. “What is Text Analysis, Really?”, *Literary and Linguistic Computing* 18, no. 2 (2003): 209-219, https://doi.org/10.1093/llc/18.2.209.
- Wiedemann, Gregor. “Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences.”. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14, no 2 (2013), DOI: http://dx.doi.org/10.17169/fqs-14.2.1949.