



# Developing a platform for the analysis of historical sources... and more

Sara Tonelli Digital Humanities Joint Research Project Fondazione Bruno Kessler



# The Digital Humanities joint research project



ALESSANDRO MARCHETTI E-mail: amarchetti@fbk.eu PHONE: 0461314974

VIEW PROFILE



STEFANO MENINI E-mail: menini@celct.it PHONE: 0461314942

```
VIEW PROFILE
```



GIOVANNI MORETTI E-mail: moretti@fbk.eu PHONE: 0461314807

VIEW PROFILE



RACHELE SPRUGNOLI E-mail: sprugnoli@fbk.eu PHONE: 0461314879



SARA TONELLI E-mail: satonelli@fbk.eu PHONE: 0461314542 VIEW PROFILE



EMANUELE PIANTA

VIEW PROFILE





•

*Graphs, Maps, Trees: Abstract Models for a Literary History*, Franco Moretti, 2007, Verso, London.

Mode of analysis called *distant reading* (vs. *close reading*), where thousands of books are automatically scanned and analyzed in an attempt to understand longterm and large-scale trends

Empirical data are the key to better knowledge of literature because they are "independent from interpretation"



•

•

•

Analysis of Language and Content In a Digital Environment:

- A digital history project: *"capacity, accessibility, flexibility, diversity, manipulability, interactivity and*
- hypertextuality" (Cohen and Rosenzweig, 2005).
- First collaboration between ICT center and Italian-German Historical Institute at FBK
- **Goal**: develop an online platform to perform temporal, geographical, linguistic and semantic analysis of historical documents
- Tailor state-of-the-art tools for Natural Language Processing to the requirements of researchers in contemporary history



•

•

•

#### Requirements:

- Select the temporal span of the research on the fly
- Pass easily from *distant* to close reading
- The platform should not be bound to a specific historical period
- Reduce manual effort to the minimum
- Intuitive interface, quick system response, modularity Multilinguality
  - Standard formats (XML for input files, HTML5 for visualization)



## ALCIDE Project: : First use case

 Complete collection of Alcide De Gasperi's writings (1881 – 1954)
 Around 3.000 documents, 3.000.000 tokens (~70.000 lemmas)
 Includes articles, speeches transcriptions, official documents, etc. Akide De Gaperi Scritti e discorsi politici



Volume IV. Tonio 1.

Alcide De Gasperi e la stabilizzazione della Repubblica 1948-1954

il Mulino



CORPUS

FILES

DeGasperi

陆 Davanti ad una bara

Ebbene, parliamone!

🖺 Il grandioso convegno di Ala

E se non ridi... che rider suoli?

Dov'erano?

Fate anche voi

La voce del popolo

🕒 Dal fanatismo alla perfidia

Il signor Capocomune al Parlamento

Verbale della seduta del Consiglio comunale

Verbale della seduta del Consiglio comunale

Elezioni dietali per la designazione delle candidature del partito popolare nei collegi rurali

# ALCIDE: "Distant" reading

1. Documents distribution over time (metadata: **year** of publication)

#### DeGasperi



Time bar to select period of interest

Documents list for the selected year

**Q:** In which year did De Gasperi publish most?



#### 2. Documents distribution per **place** of publication



**Q**: In which city were most documents published during World War I?



#### 2. Documents distribution per **place** of publication



**Q**: In which city were most documents published after World War II?



#### 3. Lemmas & tokens distribution over time Pre-processing with PoS tagger and lemmatizer from TextPro



**Q**: When did the issue of refugees become relevant for De Gasperi?



#### 3. Lemmas & tokens distribution over time Pre-processing with PoS tagger and lemmatizer from TextPro



**Q:** When was "world war" first mentioned by De Gasperi?



4. Search & visualization of key-concepts

DeGasperi

FONDAZIONE BRUNO KESSLER



Key-concepts extracted at document level and then merged

Key-concept suggestion

**Q:** Which were the main topics dealt with by De Gasperi between 1914 and 1918?



#### 5. Search and visualization of persons' names

Palmiro Togliatti	Persons
Fu uno dei membri fondatori del Partito Comunista d'Italia e, dal 1927 fino alla morte, segretario e capo indiscusso del Partito Comunista Italiano, del quale era stato il rappresentante all'interno del Comintern (di qui, per le sue capacità di mediatore fra le varie anime del partito, lo pseudonimo di "-giurista del Comintern-" attribuitogli da Lav TrotskyGiorgio Bocca, Togliatti: Capitolo VI; La scoperta della Russia - La lettera di Gramsci. Pagina 133. Arnoldo Mondadori Editore S.p.A. per La Biblioteca di Repubblica, 2005.), l'organizzazione internazionale dei partiti comunisti. Anche di questo organismo Togliatti fu uno degli esponenti più rappresentativi e, dopo che esso fu sciolto nel 1943 e sostituto dal Cominform nel 1947, rifiutò la carica di segretario generale, offertagli direttamente da Stalin, preferendo restare alla testa del partito in Italia. Dal 1944 al 1945 ricopri la carica di vice Presidente del Consiglio e dal 1945 al 1948 quella di Ministro di Grazia o Giustizia nei coverni che ressaro l'Italia dopo la caduta del fascismo. Membro dell'Assemblea Costituente, dopo le	Palmiro Togliatti       698         Pietro Nenni       618         Gesù Cristo       428         Benito Mussolini       428         Benito Mussolini       428         Giolitti       184         Leone XIII       153         Cesare Battisti       142         Iosif Stalin       139         Dio       134         Bonomi       130         Nitti       122         Enrico Conci       112         Lussu       110         Silvio Flo       106         Mauro Scoccimarro       105         0       250       500       750       1,000       1,750       2,000       2,250       2,500         Person Frequency
Entity Distrib	oution
0.15	Palmiro Togliatti
-0.05 1901 ' 1904 ' 1907 ' 1910 ' 1913 ' 1916 ' 1919 ' 1922 ' 1925 ' 1929	1932 1935 1938 1941 1944 1947 1950 1953
Drag tags here	Tags 💊
rom 1901 - To 1954 Distribution Geographical Show Terms Frequency Keywords Full Text Search Timeline Persons	<b>Q:</b> Which person is most frequent



6. Timeline personalization, visualization and comparison



**Q**: What speeches did De Gasperi give during the Russian revolution?



### ALCIDE: "Close" reading

#### 7. Visualization at document level

#### Graph 📹

#### Verso la rigenerazione?

#### Trento, 1914-11-5

Prendendo spunto dalle parole di uno scrittore francese, De Gasperi sottolinea che l'unico effetto positivo causato dalle terribili esperienze della guerra può essere una rigenerazione morale e una ritrovata pace.

Uno dei più celebri scrittori francesi, tristemente noto un tempo per una produzione letteraria tutt'altro che buona, scrive in questi giorni nei Figaro. «Alcuni prevedono, dopo questo periodo di tensione e di sforzi, di economie, di tristezze e di emozioni gravi e dolorose, lo sfrenarsi di un incredibile ardore a divertirsi e godere la vita. Ahimé! Il lutto sarà molto diffuso negli abiti di questo inverno e la festa urterebbe troppo spesso il dolore. Ma pol, perché non si dovrebbe godere della vita in guisa semplice e buona? Godere della vita vuoi forse dire gettare il denaro dalle finestre, pagare le cose dieci volte li loro valore, essere degli snob, seguire una moda vertiginosa che cambia tutte le settimane, divertirsi a delle inezie? Da parte mia lo credo invece fermamente che dopo la guerra assisteremo ad una rigenerazione, ad un rinnovamento mirabili. Vedremo allora una Francia plena di grazia e di bellezza. In cui le città cesseranno di congestionarsi e le campagne si ripopoleranno. In cui la vita regionale ritornerà in flore, e i giovani saranno lettere e le arti ritroveranno le loro classiche ilnee: soprattutto vedremo una Francia laboriosa e caritatevole in cui nessuno morrà di fame Se fosse altrimenti, se dovessimo ritrovare ancora l'alcoolismo, la miseria, i cappelli da mille lire l'uno, il tango, gli spettacoli ignobili, l'intolleranza, la persecuzione. l'arrivismo, i processi scandalosi e le scandalose assoluzioni, allora tutti quelli che furono i soldati della grande guerra avrebbero diritto di dire, in loro nome ed in nome dei morti: Non è per questo che ci siamo battuti!». Queste parole di Maurizio Donnay, che afferma la speranza in una rinascita morale della Francia, e che si augura come epilogo della lunga e dolorosa guerra non tanto il trionfo della forza materiale, un allargamento delle frontiere, una conquista di nuove e ricche colonie e l'umiliazione degli avversari definitivamente abbattuti, quanto soprattutto un riflorimento della vita morale, una restaurazione degli spiriti e delle coscienze, un ritmo di vita più nobile e più degna, meno pervasa da crisi febbrili, più caima e ordinat più sana e schietta, meno avida di beni materiali e più giolosa dei suo morale perfezionamento, sono veramente parole di bellezza e di bontà, sono un eloquente esemplo di superiorità etica. E come in Francia il Donnay, così in inghilterra, così in Germania, così in tutti i paesi belligeranti tutti gii spiriti generosi dovrebbero purificare il loro pensiero, adergerio al di sopra del presente turbine di foilla e d'odio (flagellum iracundiae come ha definito con una parola scultorea la presente guerra Benedetto XV) nel pensiero di uno sforzo più degno, di una mèta più alta che non sia la vittoria dei nemico, la distruzione dell'awersario, lo spargimento dei terrore e della strage: la vittoria di sé stessi, li proprio perfezionamento, la salutare opera dei bene e della verità. Considerare la guerra non come lo sfogo di odil e di livori lungamente repressi, ma come una prova dolorosa da sopportare con fermezza e con abnegazione, trasformare i sacrifici e i pentimenti ch'essa impone in uno strumento di elevazione e di rigenerazione significa auspicare nella pace tempi migliori e sorti più elette, significa affrettare inconsapevolmente l'ora augurale della pace. Polché se in guerra è troppo spesso lo scatenamento di istinti belluini, tutto ciò che migliora la nostra umanità, tutto ciò che ci fa meglio intendere le voci della pietà e della bontà umana non può non essere una parola pacifica.



Hohcharts.

document level



# ALCIDE: Tag manager for documents annotation

#### 8. Manual tagging of single or sets of documents

DROP FILES HERE Member of Political Party X Prop	baganda 🗶 🕒 Writings
La Cassa centrale cattolica di mutuo soccorso Writings X	Speeches
La cultura presente e la riscossa cristiana. Discorso dello studente di filol. Alc. Degasperi al Congresso di Mazzocorona     I commenti nei circoli studenteschi – l'adunanza di ieri.	<ul> <li>Reviews</li> <li>Propaganda</li> <li>Official Docs</li> </ul>
Fetch Al 🚍 Remove 🍵	Articles
	Member of Parliament
	Member of Political Party
	Prime Minister
	S Reviewer
	🗣 Journalist
	DROP FILES HERE   La Cassa centrale cattolica di mutuo soccorso   La cutura presente e la riscossa cristiana. Discorso delo   studente di Rol. Alc. Degasperi al Congresso di Mazzocorona   I commenti nei circoli studenteschi – l'adunanza di ieri.



#### From paper to bits





•

•

•

Focus group with expert users and evaluation based on System Usability Scale questionnaire:

- Some functionalities (*Timeline*, entities linked to *Wikipedia*) should be part of a learners' view
  - Revise frequency normalization criteria
  - Relevance of key-concepts is too "subjective"
  - An automatic analysis of semantic change would be useful, for instance looking at co-occurrences: which association measures?
  - Polarity at *concept* level would be more useful than at *document* level



•

•

- How difficult is it to annotate polarity at concept level on historical data?
  - Manual annotation of "*trade union*" e "*trade unionism*", 525 sentences
  - Annotation performed using the CrowdFlower platform
  - 5 judgements for each occurrence, 4 possible labels: positive, negative, neutral, don't know.
  - Gold standard of 60 sentences annotated in-house by two interns
  - Compare crowdsourced annotation with gold standard and with automatic polarity labelling via SentiWordNet/ WordNetAffect



	Crowdsourcing	SentiWNW/ WNAffect
Total	68.30%	43%
Negative	55.50%	22%
Positive	80%	31%
Neutral	46.60%	86%
Don't know	0	n.a.

Accuracy w.r.t. gold data

Low agreement between annotators, both in-house and via crowdsourcing



- Extend the system with the functionalities highlighted during the experts' evaluation
- Tag the corpus to distinguish between transcriptions of speeches, official documents, propaganda writings, etc.: how do vocabulary, discourse structure, style, readability change?
- · Add other corpora, also in English
  - Use the platform to discover something relevant to history scholars!



# Conclusions

- Many lessons learnt in the first project months:
  - A platform like A.L.C.I.D.E. can be effectively used only if "digital" and "humanities" collaborate from the very beginning, must be a two-way communication
  - NLP technology must be re-thought: make processes more intuitive, include users in the loop, importance of visualization/interfaces
  - We are not just technology providers!
- Many challenges ahead:
  - New project involving MART and MUSEION for exploration platform of verbo-visual art: recommend artworks based on similarity? Crowdsourcing text transcription? How to encode the graphical component?
  - Horizon 2020