Multilingual Access to European Cultural Heritage
The Role of Human Language Technology

Alessandro Lenci

CoLing Lab – Laboratorio di Linguistica Computazionale

Dipartimento di Filologia, Letteratura e Linguistica

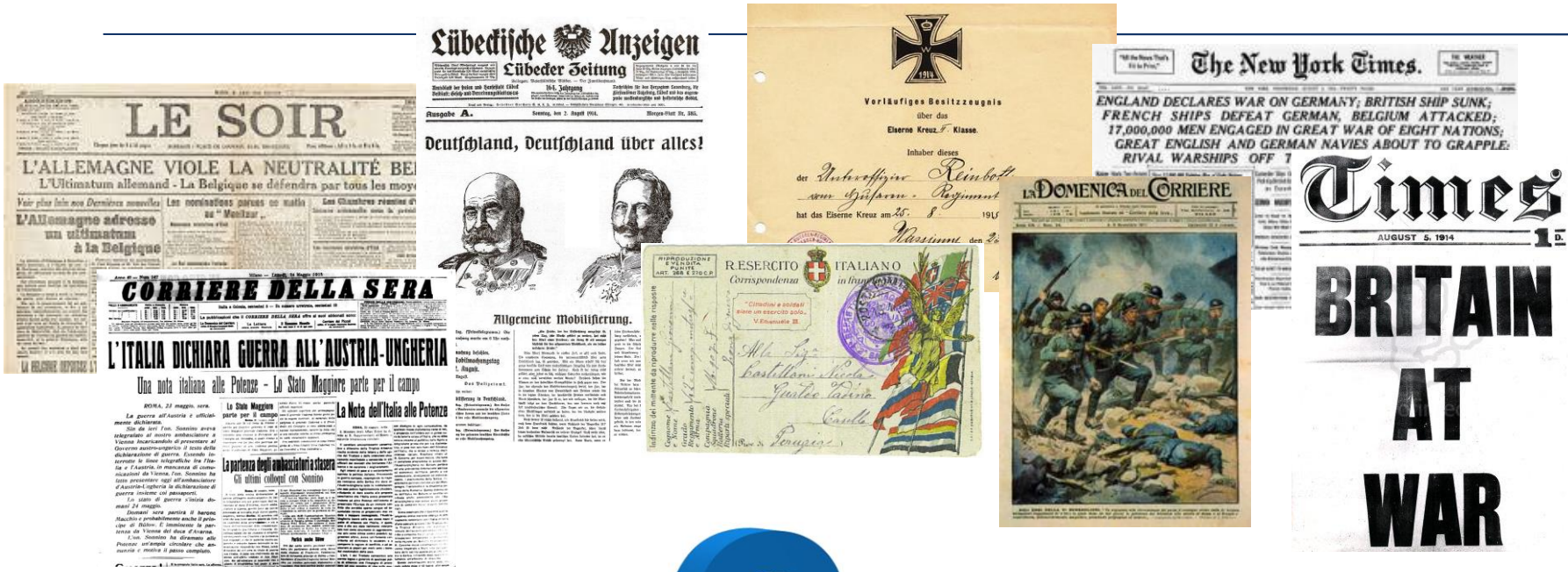
Università di Pisa

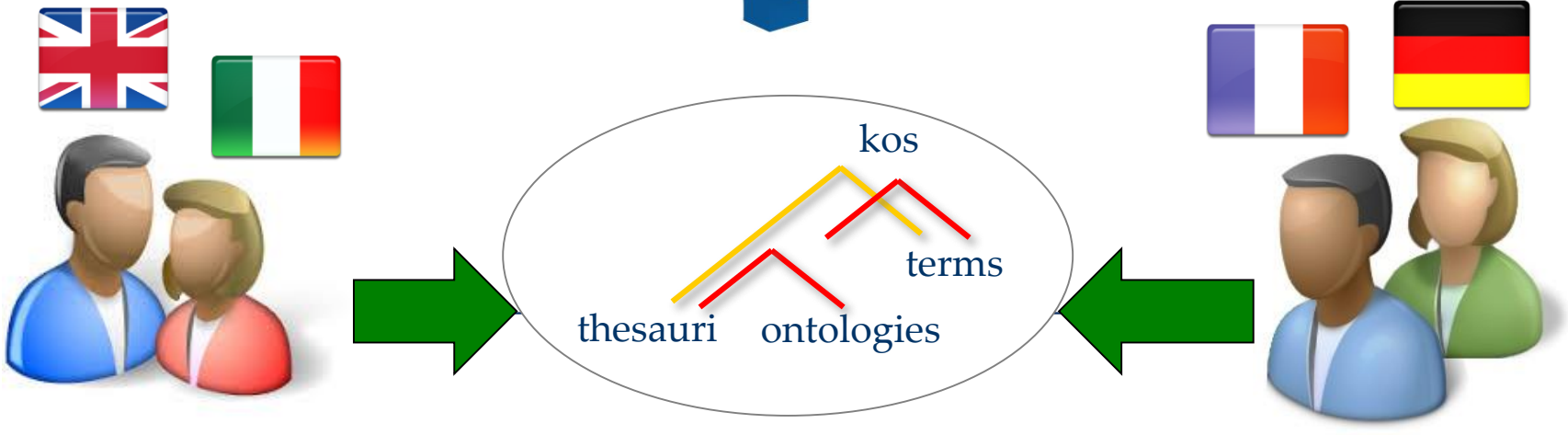
Athena Plus International Conference

Roma, 20 ottobre 2015

Multilingual Access to Multilingual Cultural Heritage

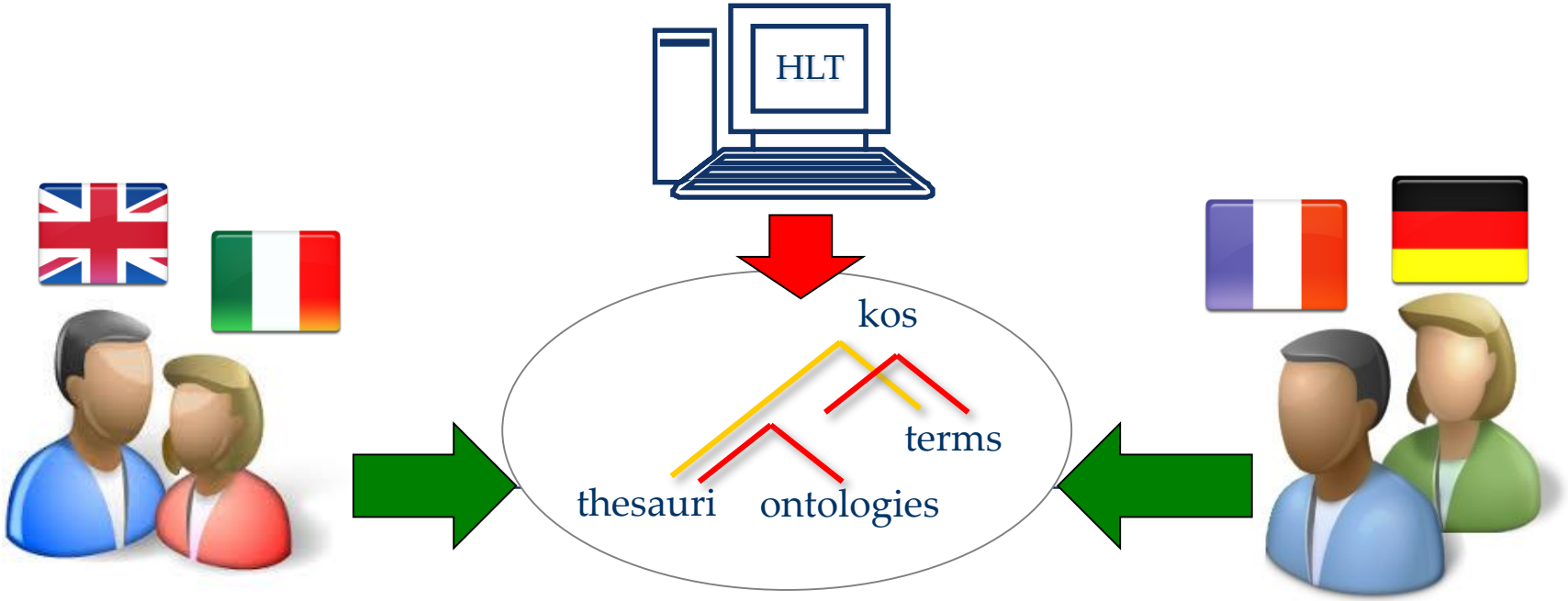
Multilingual Access to Multilingual Cultural Heritage





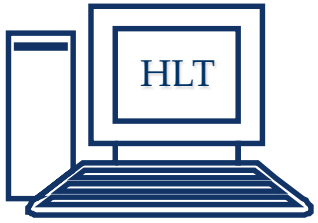
Multilingual Access to Multilingual Cultural Heritage

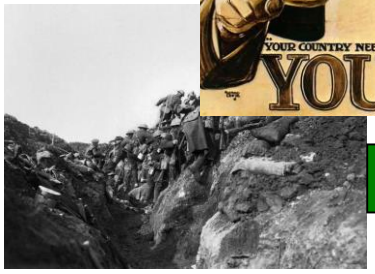
Natural Language Processing

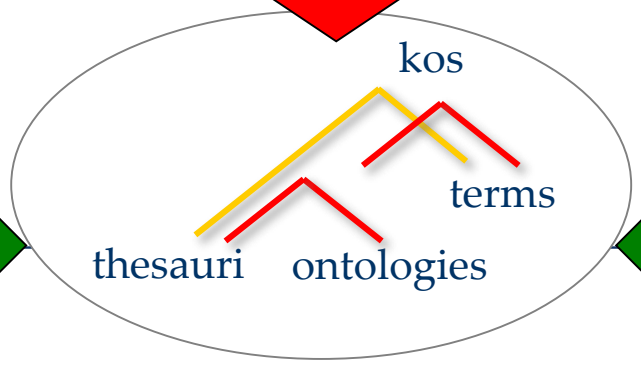


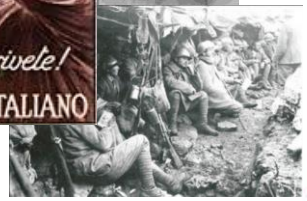
Linking Multilingual and Multimedial Cultural Heritage

Natural Language Processing







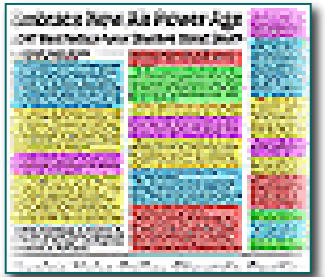


Human Language Technologies (HLT)

automatic analysis of linguistic structure

texts

Natural Language Processing



linguistically annotated texts

HLT

text indexing with semantic metadata

term extraction, and ontology learning to support the process of KOS development

kos

terms

thesauri

ontologies

War Memories (Memorie di Guerra)
http://www.memoriediguerra.it/wwm/

- An ongoing project to carry out a computational analysis and semantic indexing of Italian texts about WWI and WWII
  - University of Pisa, CoLing Lab
  - ILC-CNR, Pisa
  - history consultant: Prof. Nicola Labanca (University of Siena)
- Texts are annotated automatically with state-of-the-art NLP tools to extract various kinds of information
  - simple and multi-word terms
  - named entities
  - events and their participants
  - georeferenced locations

A first Application: Italian War Bulletins

- Issued by the Italian *Comando Supremo* "Supreme Headquarters" during WWI and WWII as the official daily report about military operations
  - WWI: 1,342 texts from 24 May 1915 to 11 November 1918
    - » published in 1923, never digitalized before (189,783 tokens)
  - WWII: 1,201 texts from 10 June 1940 to 8 September 1943
    - » published in 1970, available in html (211,854 tokens)

1° *luglio.*

Nella zona del Tonale le nostre artiglierie aprirono il fuoco sulle posizioni di Monticello e di Saccarana, disperdendovi reparti nemici intenti a lavori di apprestamenti e difesa.

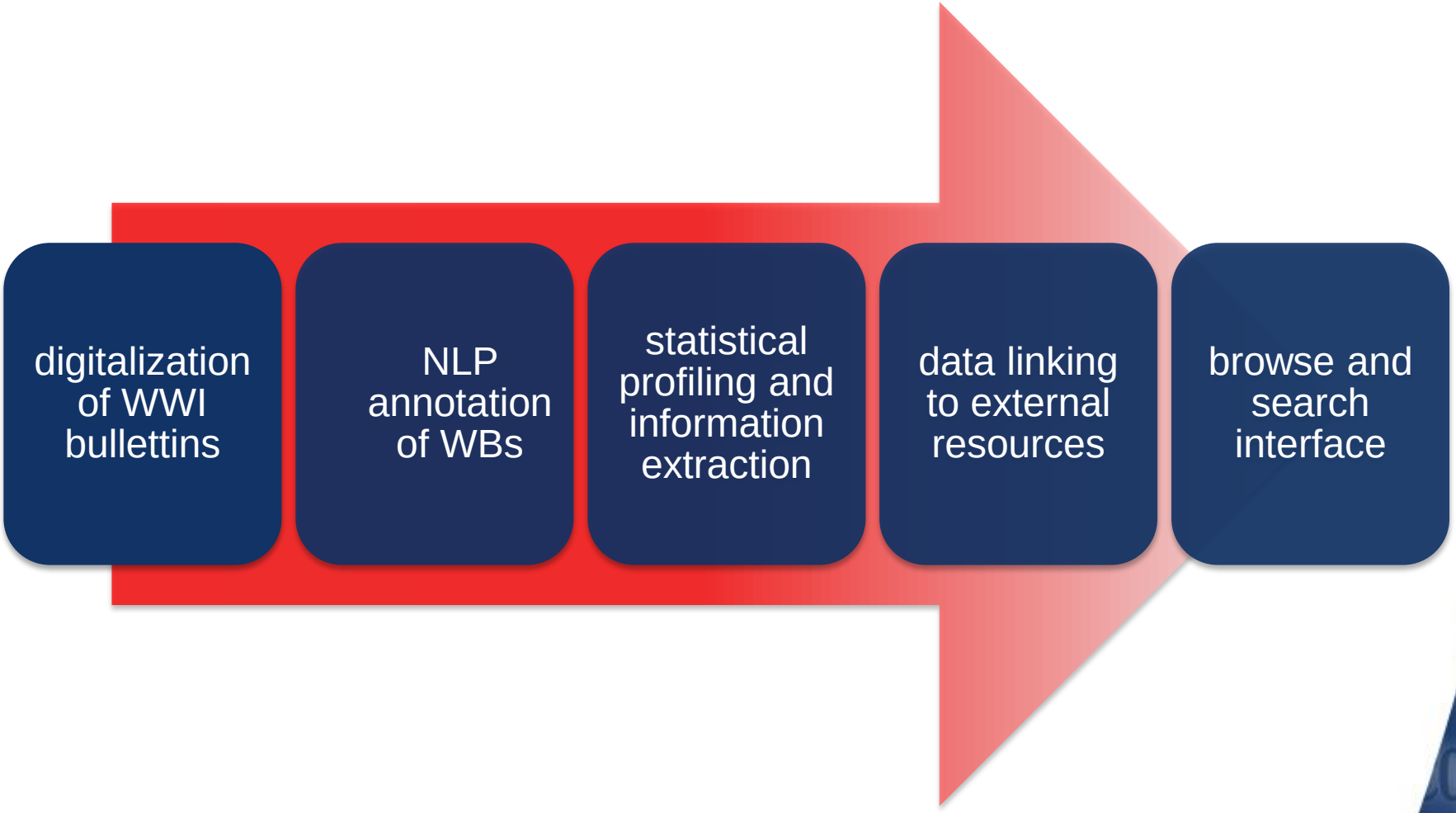
In Val Padola pattuglie di ufficiali arditamente spinte sul Seikofl vi accertarono la costruzione, per parte del nemico, di trinceramenti con reticolati, che la nostra artiglieria battè poi con efficacia.

In Carnia il nemico ha tentato vigorosi attacchi notturni contro le nostre posizioni del Passo di Monte Croce e del Pal Piccolo, aiutandosi con razzi e riflettori e lanciando bombe contenenti gas asfissianti. Fu in entrambi i punti respinto. Disperdemmo, mediante tiri di artiglieria, nuclei di lavoratori apparsi sulle pendici settentrionali del Freikofel e del Pal Grande e lungo la mulattiera di Val Bombasch.

Fu ripreso con buoni risultati il tiro sul forte Hensel.

Alla testata di Valle Resia l'importante posizione di Banjski Skedenj, dominante la conca di Plezzo, venne da noi solidamente occupata.

Building War Memories



Digitalization of WWI Bulletins

- OCR was performed with Tesseract
- Manual revision and annotation with XML metadata
  - future release of texts in TEI-XML

```
<doc url="http://www.ilc.cnr.it/w2m/doc49.html" index="49" day="39" date="1 luglio 1915">
<date>1 luglio.</date>
<p>Nella zona del Tonale le nostre artiglierie aprirono il fuoco
sulle posizioni di Monticello e di Saccarana, disperdendovi reparti nemici intenti a lavori di apprestamenti e difesa.</p>
<p>In Val Padola pattuglie di ufficiali arditamente spinte sul Seikofl vi accertarono la costruzione, per parte del nemico, di trinceramenti con reticolati, che la nostra artiglieria batté poi con efficacia.</p>
<p>In Carnia il nemico ha tentato vigorosi attacchi notturni contro le nostre posizioni del Passo di Monte Croce e del Pal Piccolo, aiutandosi con razzi e riflettori e lanciando bombe contenenti gas asfissianti. Fu in entrambi i punti respinto. Disperdemmo, mediante tiri di artiglieria, nuclei di lavoratori apparsi sulle pendici settentrionali del Freikofel e del Pal Grande e lungo la mulattiera di Val Bombasch.</p>
<p>Fu ripreso con buoni risultati il tiro sul forte Hensel.</p>
<p>Alla testata di Valle Resia l'importante posizione di Banjski
Skedenj, dominante la conca di Plezzo, venne da noi solidamente occupata.</p>
```

Text Processing with NLP Tools

<doc url="http://www.ilc.cnr.it/w2m/doc49.html" index="49" day="39" date="1 luglio 1915">

1	Nella	in	E	EA	num=s|gen=f	8	comp	0
2	zona	zona	S	S	num=s|gen=f	1	prep	0
3	del	di	E	EA	num=s|gen=m	2	comp	0
4	Tonale	Tonale	S	SP	_	3	prep	B-LOC
5	le	il	R	RD	num=p|gen=f	7	det	0
6	nostre	nostro	A	AP	num=p|gen=f	7	mod	0
7	artiglierie	artiglieria	S	S	num=p|gen=f	8	subj	0
8	aprirono	aprire	V	V	num=p|per=3|mod=i|ten=s	0	ROOT	0
9	il	il	R	RD	num=s|gen=m	10	det	0
10	fuoco	fuoco	S	S	num=s|gen=m	8	obj	0
11	sulle	su	E	EA	num=p|gen=f	8	comp	0
12	posizioni	posizione	S	S	num=p|gen=f	11	prep	0

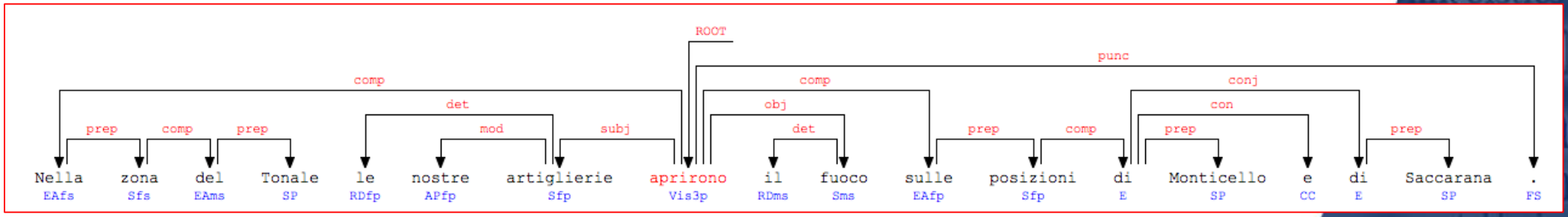
tokenization

lemmatization

PoS tagging

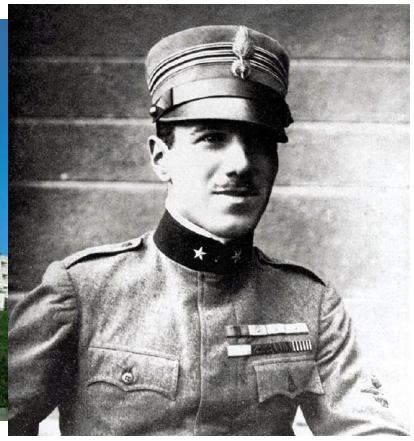
dependency parsing

named entity recognition



Named Entity Recognition

- Automatic identification and semantic classification of named entities in texts
  - locations (LOC)
    - » *Montello*
  - persons (PER)
    - » *Francesco Baracca*
  - military units (MIL)
    - » *Brigata Sassari*
  - ships (SHP)
    - » *Czepel*
  - planes (PLN)
    - » *Aviatik*



Automatic Term Extraction

- Single and multi-word terms are automatically extracted with T2K^2 (Dell'Orletta et al. 2014) and EXTra (Passaro & Lenci 2015)

domain term
mitragliare, "to machine-gun"
spezzonare "to bomb with incendiary devices"
bombardare "to bomb"
abbattere "to shoot down"
silurare "to torpedo"
incendiare "to set on fire"
affondare "to sink"
attaccare "to attack"

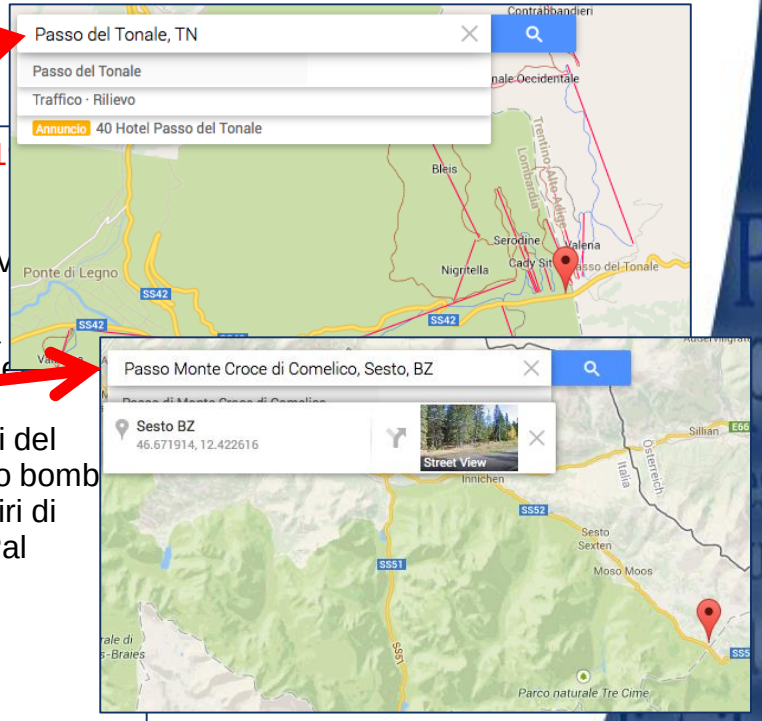
Applications:
- term-based indexing and search of texts
- semi-automatic population of thesauri and ontologies

term	LMI
fronte greco "Greek frront"	927.30
tenente di vascello "lieutenant"	699.04
lieve danno "small damage"	659.14
aereo nemico "enemy plane"	623.10
capitano di corvetta "captain"	593.89
artiglieria contraerea "flak"	548.13
bomba di grosso calibro "large bomb"	500.12
velivolo nemico "enemy plane"	496.32
bollettino odierno "today bulletin"	456.01
caccia germanico "German fighter"	441.91
obiettivo militare "military target"	423.78
campo di aviazione "aviation field"	422.07
vasto incendio "large fire"	416.86
caccia tedesco "German fighter"	413.63
piroscafo di medio tonnellaggio "average ton ship"	366.60

Data Linking

- Location NEs have been georeferenced semi-automatically
  - particularly challenging because of spelling variations (e.g., arabic names in WWII African campaigns), historical changes, etc.

<doc url="http://www.ilc.cnr.it/w2m/doc49.html" index="49" day="39" date="1 luglio 1
<date>1 luglio.</date>
Nella zona del **Tonale** le nostre artiglierie aprirono il fuoco
sulle posizioni di Monticello e di Saccarana, disperdendovi reparti nemici intenti a lav
apprestamenti e difesa.
In Val Padola pattuglie di ufficiali arditamente spinte sul Seikofl vi accertarono la
costruzione, per parte del nemico, di trinceramenti con reticolati, che la nostra artiglie
batté poi con efficacia.
In Carnia il nemico ha tentato vigorosi attacchi notturni contro le nostre posizioni del
Passo di Monte Croce e del Pal Piccolo, aiutandosi con razzi e riflettori e lanciando bomb
contenenti gas asfissianti. Fu in entrambi i punti respinto. Disperdemmo, mediante tiri di
artiglieria, nuclei di lavoratori apparsi sulle pendici settentrionali del Freikofel e del Pal
Grande e lungo la mulattiera di Val Bombasch.
Fu ripreso con buoni risultati il tiro sul forte Hensel.
Alla testata di Valle Resia l'importante posizione di Banjski
Skedenj, dominante la conca di Plezzo, venne da noi solidamente occupata.



Conclusions and Outlook

- CH texts are deemed to be hard for HLT…
  - digitization errors, diachronic variations, substandard forms, etc.
- …, but "noisy input" is a common challenge for HLT
  - cf. speech processing, social media analysis, etc.
- HLT are not error free, but speed up the process of KOS creation and enhance CH documents with semantic metadata

- Computational linguistics and HLT offer important perspectives to manage and exploit cultural heritage
  - semantic searches
  - multimedia resources crosslinking
  - multilingual access to information content

Thank You!

Questions?

UNIVERSITY OF PISA

COMPUTATIONAL LINGUISTICS LAB

http://colinglab.humnet.unipi.it