A Journey in Linguistic Computing from Father Busa to Linguistic Linked Data

Marco Passarotti

Lectio Conference

Imagining the Future of Pre-Modern Intellectual History

KU Leuven - June 2nd, 2021

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.

Overview

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

Table of Contents

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

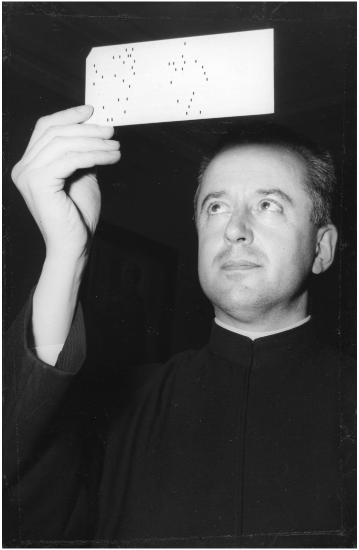
The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

Father Roberto Busa

The impossible takes a little longer



Father Roberto Busa

Punching Cards ...until the magnetic tapes came

A Nightmare intervened

Busa, R. *L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione*. 1962.

[…] a nightmare intervened, technology triumphant with its latest creation: automation. People shuddered, considering it a crude, hard bulldozer that goes roaring ahead, crushing and shredding flowers, amongst which, **a delicate and gentle victim**, is humanism. […]

A Nightmare intervened

Busa, R. *L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione*. 1962.

[…] a nightmare intervened, technology triumphant with its latest creation: automation. People shuddered, considering it a crude, hard bulldozer that goes roaring ahead, crushing and shredding flowers, amongst which, **a delicate and gentle victim**, is humanism. […]

What happened was sensational: a machine made us realize that no humanist is such command of his own language as to be able to answer such questions.

A machine […] has revealed that **there is still too little humanism of the serious and systematic type**.

A Nightmare intervened

Busa, R. *L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione*. 1962.

[…] a nightmare intervened, technology triumphant with its latest creation: automation. People shuddered, considering it a crude, hard bulldozer that goes roaring ahead, crushing and shredding flowers, amongst which, **a delicate and gentle victim**, is humanism. […]
What happened was sensational: a machine made us realize that no humanist is such command of his own language as to be able to answer such questions.
A machine […] has revealed that **there is still too little humanism of the serious and systematic type**.
Economic facts today demand a qualitative increase of grammatical and lexical sciences as one of the necessary conditions of their vital development.

English translation from: Nyhan Julianne, Passarotti Marco (eds.), *One Origin of Digital Humanities. Fr. Roberto Busa in His Own Words*. Cham, Springer International Publishing, 2019.

Table of Contents

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

Rules and Constituency Grammars

Intuition-based NLP

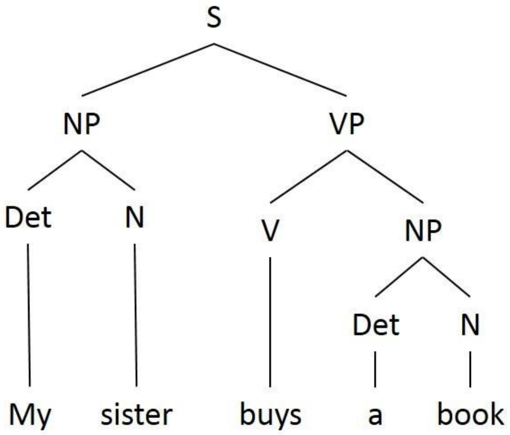


Table of Contents

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

Linguistic Resources and NLP

There is nothing better than more (meta)data ...really?

- Since the 90s: more Lexical and Textual Resources for more languages
- The more (Meta)data you have, the more (Meta)data you build: training data-driven NLP tools

Table of Contents

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

The CLARIN Infrastructure

One place fits all

Table of Contents

(Pre)historical Steps

- Father Busa: Punched Cards and Magnetic Tapes
- Rules and Constituency Grammars

The Empirical Revolution and the WWW

- Linguistic Resources and Machine Learning
- Infrastructures: Making Resources Findable, Accessible and Reusable
- Linguistic Linked Open Data: Making Resources Interoperable

Conclusion: Imagining (and hoping for) the future

Research question

State of affairs

We have built and collected (for Latin and other languages):

Research question

State of affairs

We have built and collected (for Latin and other languages):

- Textual Resources

Research question

State of affairs

We have built and collected (for Latin and other languages):

- Textual Resources
- Lexical Resources

Research question

State of affairs

We have built and collected (for Latin and other languages):

- Textual Resources
- Lexical Resources
- NLP Tools

Research question

State of affairs

We have built and collected (for Latin and other languages):

- Textual Resources
- Lexical Resources
- NLP Tools

Scattered and unconnected

The Linked Data Principles

…just to be FAIR

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)

The Linked Data Principles

...just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL

The Linked Data Principles

…just to be FAIR

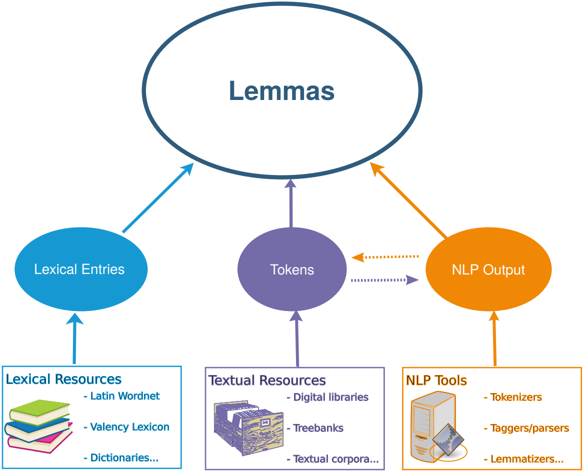
- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL
- Include links to other URIs

The LiLa Knowledge Base

One place interlinks all …at least for Latin

LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



Lemmas

Lexical Entries

Tokens

NLP Output

Lexical Resources
- Latin Wordnet
- Valency Lexicon
- Dictionaries...

Textual Resources
- Digital libraries
- Treebanks
- Textual corpora...

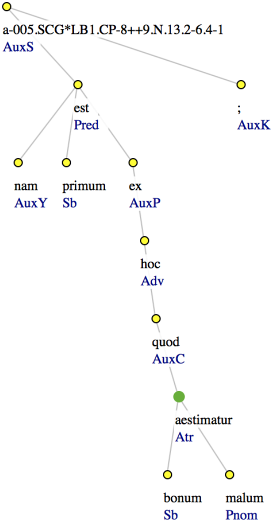
NLP Tools
- Tokenizers
- Taggers/parsers
- Lemmatizers...

From the IT Punched Cards to LiLa

Source: the *Index Thomisticus* Treebank (original scheme)

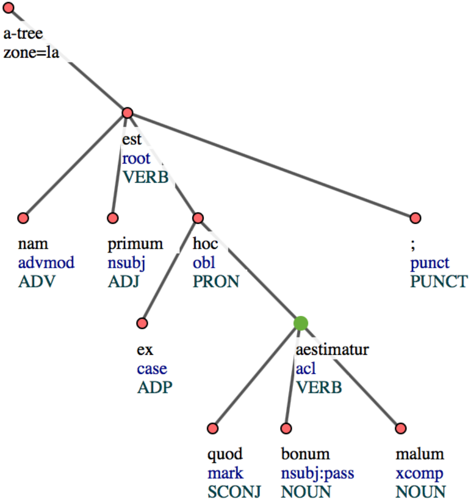
*nam primum est ex hoc quod bonum **aestimatur** malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

*for the first arises because the good **is judged** to be evil*; (Trans. Anton C. Pegis)



From the IT Punched Cards to LiLa

Source: the *Index Thomisticus* Treebank (UD scheme)



From the IT Punched Cards to LiLa

A token from the *Index Thomisticus* Treebank in LiLa

Token *aestimatur*

`https://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.2-6.4-1W8`

Challenges and Perspectives

Imagining (and Hoping for) the Future

- Skills in the Humanities are needed to transform information into knowledge. A large amount of Big Data is linguistic data

Challenges and Perspectives

Imagining (and Hoping for) the Future

- Skills in the Humanities are needed to transform information into knowledge. A large amount of Big Data is linguistic data
- A qualitative turn in quantitative analysis of linguistic data

Challenges and Perspectives

Imagining (and Hoping for) the Future

- Skills in the Humanities are needed to transform information into knowledge. A large amount of Big Data is linguistic data
- A qualitative turn in quantitative analysis of linguistic data
- To (really!) impact the world of the Humanities

Thanks!

Get in touch

LiLa: Linking Latin
Università Cattolica del Sacro Cuore
CIRCSE Research Centre

`info@lila-erc.eu`

`https://github.com/CIRCSE`

`https://lila-erc.eu`

`@ERC_LiLa`

Largo Gemelli 1, 20123 Milan, Italy

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.

Credits