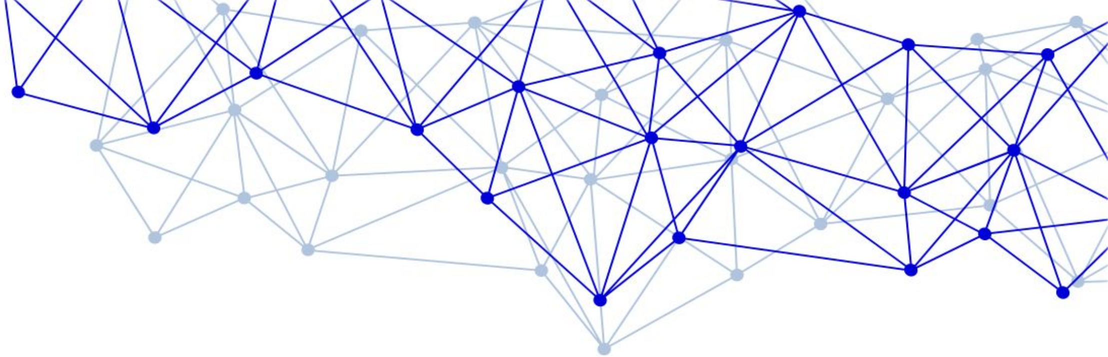
MUNI
ARTS



Lexicographic considerations in the coding of inquisition transcripts of medieval Latin

Gideon Kotzé, David Zbíral and Robert L. J. Shaw

29 June 2023

Masaryk University • Centre for the Digital Research of Religion
DISSINET – Dissident Networks Project
www.dissinet.cz

erc
European Research Council
Established by the European Commission

Acknowledgements

European Research Council

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 101000442 “Networks of Dissent: Computational Modelling of Dissident and Inquisitorial Cultures in Medieval Europe”, 9/2021–8/2026).

Centre for the Digital Research of Religion (CEDRR)

Masaryk University,
Faculty of Arts, Department
for the Study of Religions.

Introduction

- Historians working with languages of the past are required to have an expert understanding of source texts and place them in the correct context.
- Many historians of the medieval era work intensely on extracting meaning from source texts, requiring them to also have a good understanding of the languages in question.

CASTEMO

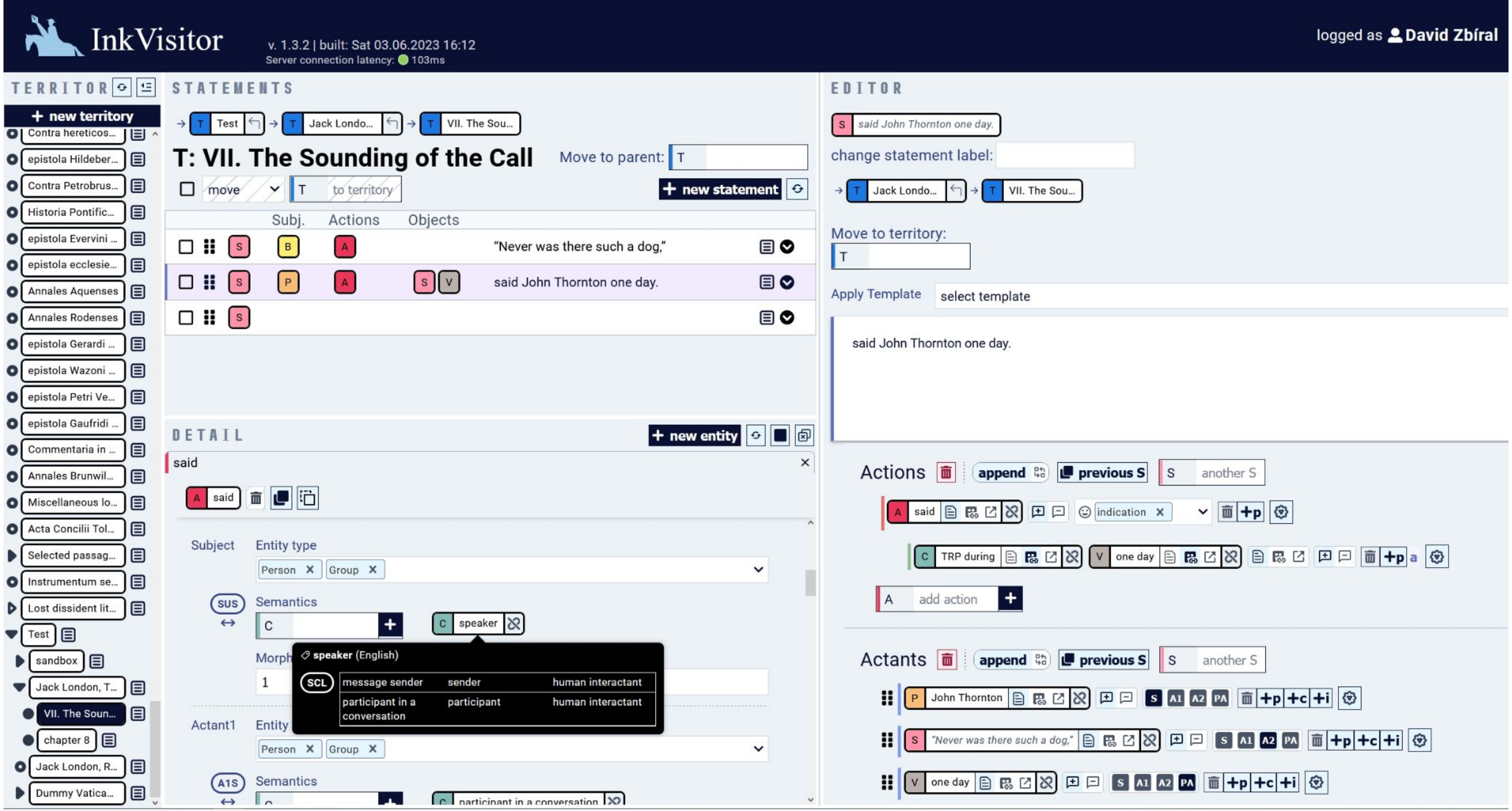
- Annotating **transcribed medieval inquisition** registers using the recently developed **Computer-Assisted Semantic Text Modelling (CASTEMO)**.
- Inspired by well-known ideas on meaning representation: RDF, OWL, Semantic Web, Quantitative Narrative Analysis
- Human-controlled, computer assisted, modeling the source closely: lexically (original language), syntactically, semantically, and contextually (part of document, context, textual order).
- Semantic units called statements that allow to model virtually everything in the source text
    - semantics, syntax, discursive elements, analytical layers (epistemic level, certainty, modality), conflicts, ambiguous evidence

CASTEMO data model

- Founded on **entities**:
  - 2 “types”: Action, Concept
  - 10 “individuals” (specific entities): Statement, Resource, Territory, Person, Group, Living Being, Object, Location, Event, Value
- **Related** in different ways:
  - **Statements** link an “action” (governed by an Action entity) to “actant” entities
  - **Properties** link a source entity to a “property value” entity via a “property type” (always a Concept); exist both within and outside Statement context
  - **Relations** are of several predefined types (e.g. Class, Superclass, Synonym, Action/Event Equivalent), which sets core semantic and ontological links

CASTEMO

- Covers statement chains (main → subordinate clause…).
- Covers modalities (e.g. question, wish, rather than just indication).
- Valency frames defined for any verb.
- More info: https://muni.cz/go/castemo.
- Open-source, browser-based data collection interface: **InkVisitor**, https://inkvisitor.net



T: VII. The Sounding of the Call

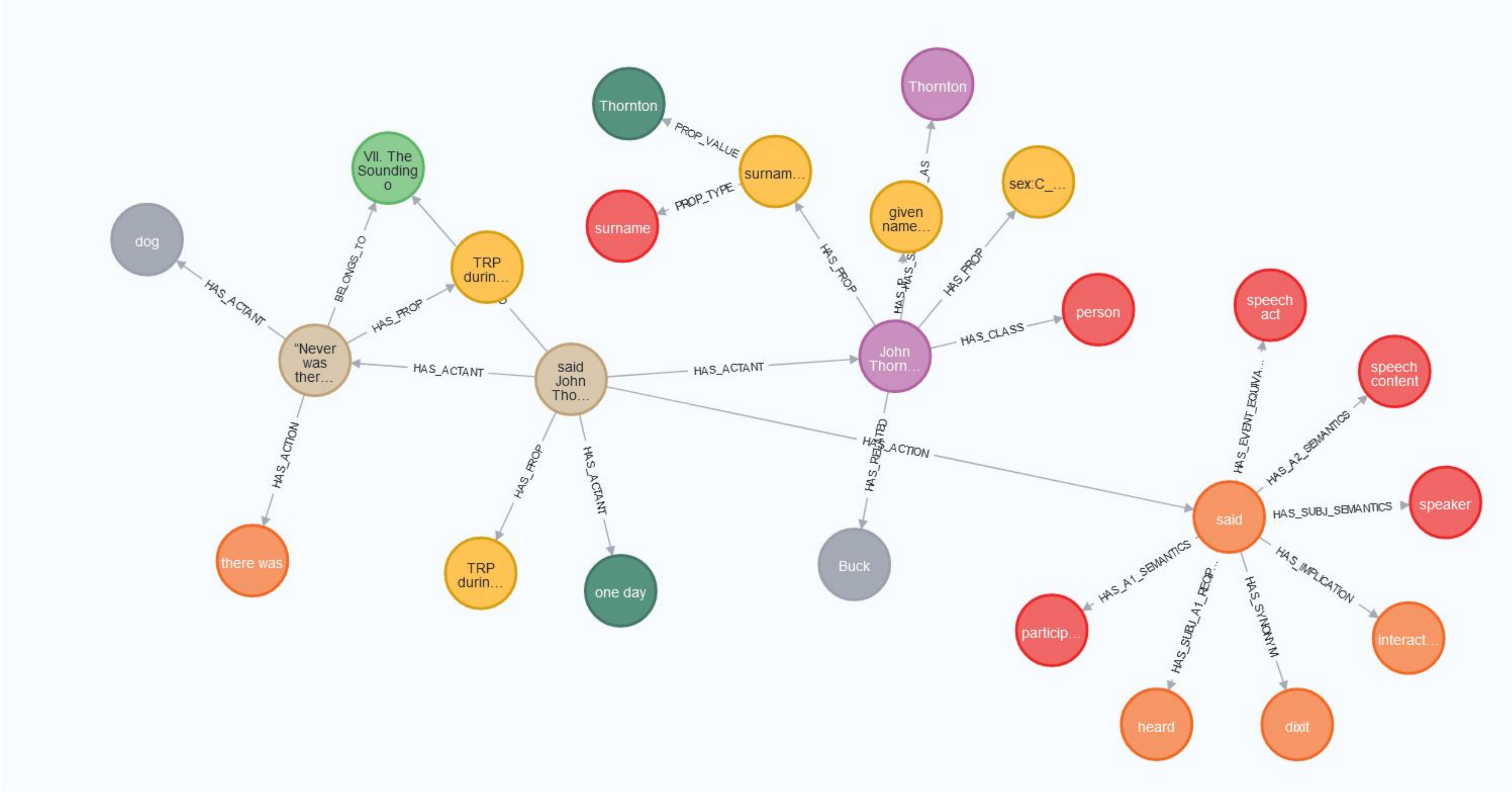
"Never was there such a dog,"

said John Thornton one day.

speaker (English)

SCL		
message sender	sender	human interactant
participant in a conversation	participant	human interactant

said John Thornton one day.


Thornton
PROP_VALUE
surnam…
PROP_TYPE
surname
Thornton
_AS
given name…
sex:C_…
HAS_PROP
HAS_PROP
VII. The Sounding o
dog
HAS_ACTANT
BELONGS_TO
HAS_PROP
TRP durin…
"Never was ther…
HAS_ACTANT
said John Tho…
HAS_ACTANT
John Thorn…
HAS_CLASS
person
speech act
speech content
HAS_ACTION
HAS_ACTION
HAS_RELATED
HAS_PROP
HAS_ACTANT
HAS_EVENT_EQUIVA…
HAS_A2_SEMANTICS
HAS_SUBJ_SEMANTICS
speaker
said
there was
TRP durin…
one day
Buck
HAS_A1_SEMANTICS
HAS_SUBJ_A1_PROP…
HAS_SYNONYM
HAS_IMPLICATION
particip…
interact…
heard
dixit

Valencies

- **Syntax**: actant slots (argument structure)
- **Semantics**: semantic roles (in a sense; restricts type of entity that may fill a slot), lexical and compositional
- **Lexical** (collocability): for example, requiring certain prepositions such as "cum", to precede a certain argument

Concepts and Actions

- Coding follows principles of **knowledge graph** creation, i.e.
  - entities, relationships, events, properties, metadata, etc. that follow a semantic data model
  - can be processed efficiently and unambiguously by a computer (database representation and software)
  - network of related data points, properties, semantic relationships
- Apart from individual entities (Persons, Events, Locations, Groups, etc.) we are building a **lexico-semantic network** of related Concepts and Actions. This and CASTEMO output can be used for querying and different varieties of quantitative analysis.

Concepts and Actions

- **Actions** are verbs or phrases represented as predicates in statements
- **Concepts** are other PoS (mostly nominals and adverbials).
- Part of speech tags, labels, descriptions, semantic relations.

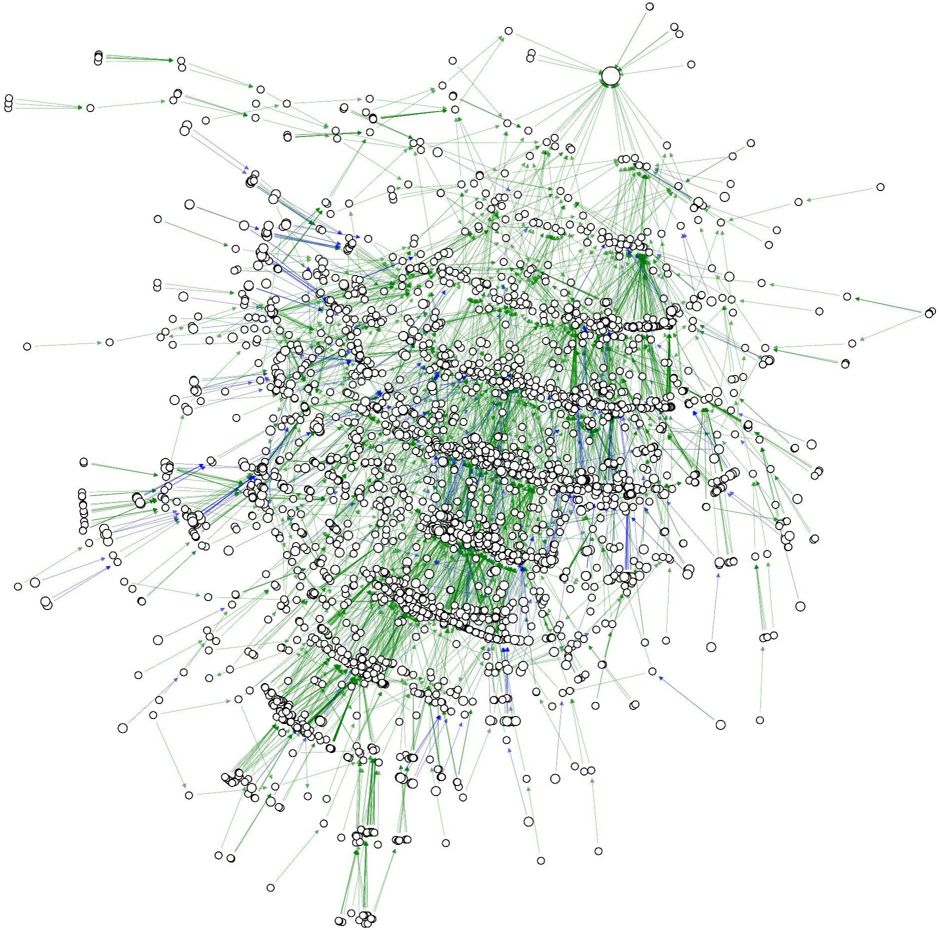
Valencies

- Any verb is a **lemma-meaning unit** (i.e. polysemes rendered as more entries).
- For any actant slot (e.g. subj), three valency types are defined
    - **Entity type valency**: what entity type this slot can take (e.g. only a person or a group).
    - **Semantic valency**: once an entity occupies this slot, it plays the role defined by the Concept (e.g. the subject of “say” is “speaker”).
    - **Morphosyntactic valency**: abbreviations defining the preposition and case for that actant; helps coders to decide whether this is the right Action (and in the near future, automatic parsers to semantically disambiguate Latin verbs on the basis of morphosyntax).

Relation name	Abbreviation	Inverse relation name	Entity combinations	Detail	Example
Superclass	SCL	Subclasses	A-A, C-C	Superordinate term (hypernym).	C apple -> C fruit A walk -> A move
Synonym	SYN	Synonym	A-A, C-C	Synonym both within a language and across languages.	C funny <-> C strange
Antonym	ANT	Antonym	A-A, C-C	Opposite term.	C good <-> C bad
Property Reciprocal	PRR	PropertyReciprocal	C-C	The concept reciprocated the other way.	C mother <-> C child
Action/Event Equivalent	AEE	Action equivalent	A-C	What is this action in the world of nouns?	A baptize -> C baptism
Holonym	HOL	Meronyms	C-C	Relation of a part to its whole.	C gate of a monastery -> C monastery
Implication	IMP	Used as Implication	A-A	Action implied by this action.	A dine (with sb) -> A be in the company (of sb)

Statements	> 11k
Entities	33,533
Actions	661
Latin Actions	552
Concepts	4,319
Latin Concepts	2,029
Relations	15,727
Relations within Concepts and Actions	6,929
Superclass Relation	3,610
Synonym Relation	412
Action/Event Equivalent Relation	450
Actant Semantics Relations	1,602

Kamada-Kawai force directed graph of Concepts and Actions with Superclass (green) and Action/Event Equivalent (blue) relations (Created: 29/04/2023)



External links and database

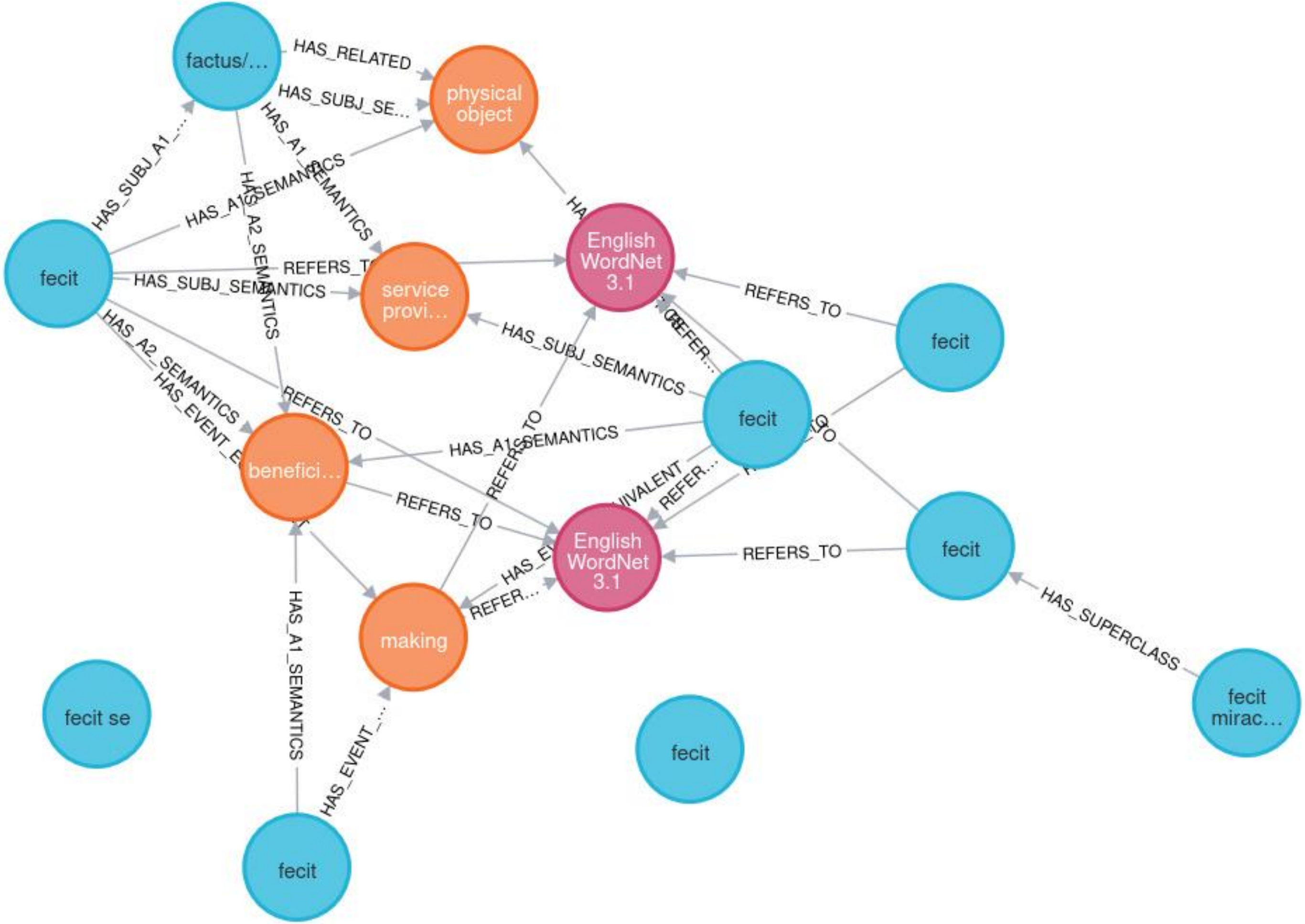
- We link to English WordNet synset IDs and/or sense keys.
- CASTEMO output (comprising a syntactic-semantic treebank) and network are stored in a **documented-oriented JSON database** (RethinkDB).
- Being compatible with a **graph-based approach**, we recently produced a **Neo4j** database projection that can be queried in a more intuitive way and for which tools exist that can produce helpful visualizations (e.g. Neo4j Browser).

Corpus

- **> 1m tokens** and growing (mid-thirteenth to early sixteenth century and from central Italy to England)
- Vast majority in Latin.
- All annotation is based on digitized transcriptions (represented in corpus).
- Thanks to alignment with the text, Statements, Concepts, Actions can be **enriched** with (1) textual context and (2) layers of linguistic analysis from NLP tools.

Refined querying

- CQ systems can apply **POS tagging** and **lemmatization** to assist with **disambiguation** of homonyms or polysemes, such as the verb *facere* ("to make, do, or accomplish; become [passive]")
- However, being able to filter a search by additional means, such as valency patterns and different types of semantic relations, would assist where we require, for example, a human subject and indirect object (Person or Group), as well as a non-human direct object (Concept or Object).
- We would be able to filter by specific **semantic relations** (e.g. it must act as a superclass for the entry "fecit miraculum" (“performed a miracle”) or must match with specific WordNet senses, etc.)


factus/…
HAS_RELATED
physical object
HAS_SUBJ_SE…
HAS_A1_SEMANTICS
HAS_SUBJ_A1_…
HAS_A1_SEMANTICS
HAS_A2_SEMANTICS
fecit
REFERS_TO
HAS_SUBJ_SEMANTICS
service provi…
English WordNet 3.1
REFERS_TO
fecit
HAS_SUBJ_SEMANTICS
REFER…
HAS_A2_SEMANTICS
HAS_EVENT_E…
REFERS_TO
fecit
HAS_A1_SEMANTICS
benefici…
REFERS_TO
REFER…
…IVALENT
REFERS_TO
English WordNet 3.1
REFERS_TO
fecit
HAS_E…
REFER…
making
HAS_SUPERCLASS
fecit se
fecit
fecit mirac…
HAS_A1_SEMANTICS
HAS_EVENT_…
fecit

Invisible lexicography?

- Primary goal is to **code** the salient aspects of the source texts in order to answer **historical questions**, not to build linguistic Latin resources. However, we see the latter being produced as a useful byproduct.
- In effect, we are building a **linguistic resource** for a relatively **underrepresented variety of Latin** (medieval) for a **specific domain** (inquisitorial registers). The fact that it may be useful for lexicography is a **useful side effect** of our efforts.
- Coding the source registers using InkVisitor involves various **linguistic decisions** for the purpose of **meaning representation** – in a sense, performing some tasks that are associated with lexicography.

Linking Latin

- Within the body of digital Latin linguistic resources, the Linking Latin (LiLa) project is a well-known current initiative.
- LiLa follows **LOD principles** to link up several different individual Latin resources in a lemma-based approach, including dictionaries and the Latin WordNet.
- We are currently exploring how we can make use of these resources to expand our knowledge base. A clear next step is to link our lemmas and meanings to corresponding URIs in the LiLa knowledge base.

Lexicography

- Our database allows for in-depth quantitative analysis using various methods, but how can this be useful for lexicography?
- One possible approach is to **annotate our corpus** with all the semantic relations available through Statements, and upload this to Sketch Engine.
- For Word Sketches, a Word Sketch grammar can make use of NLP enriched layers (including dependency relations,[1] as well as semantic relations[2]).

[1] See e.g. Horák et al., 2009. [2] See León-Araúz et al., 2016.

Boosting CASTEMO

- Problem: The coding of statements is **slow** and covers a small subset of the corpus (ca 11k Statements).
- A possible solution to this is utilizing **machine learning** to perform semantic tasks that are currently done manually in InkVisitor. Such tasks could include **semantic role labeling** (according to the CASTEMO data model) and **relation extraction**.
- **Weak supervision** has been applied successfully to areas where a small amount of labeled data exists next to a larger amount of unlabeled data, improving on pure unsupervised approaches.

Boosting CASTEMO

- **Bidirectional Encoder Representations from Transformers (BERT)** is a language modeling approach that produces contextual representations from unlabeled text that has been used to inform and improve a number of NLP tasks, sometimes by a significant amount. For Latin, Latin BERT has been applied to word sense disambiguation and semantic search, among others (Bamman and Burns, 2020).
- It might also be possible to generate certain aspects of the CASTEMO workflow in order to speed up the process. This includes pre-selecting statements by using a syntactic parser, auto-suggesting entries in the network, etc.

Conclusion

- We have presented a knowledge base for the description of historical sources for a specialized domain (inquisition registers predominantly in medieval Latin).
- Main goal is quantitative analysis of the sources using advanced computational techniques.
- However, this resource contains useful linguistic data, including syntactic and semantic descriptions concepts and actions, and annotation of a corpus as modelled statements.
- This has the potential to be exploited by lexicography, as the entries are corpus-based and/or can be verified by corpus analysis.

Děkuji mockrát / Thank you very much!

gideon.kotze@mail.muni.cz

david.zbiral@mail.muni.cz

robert.shaw@mail.muni.cz

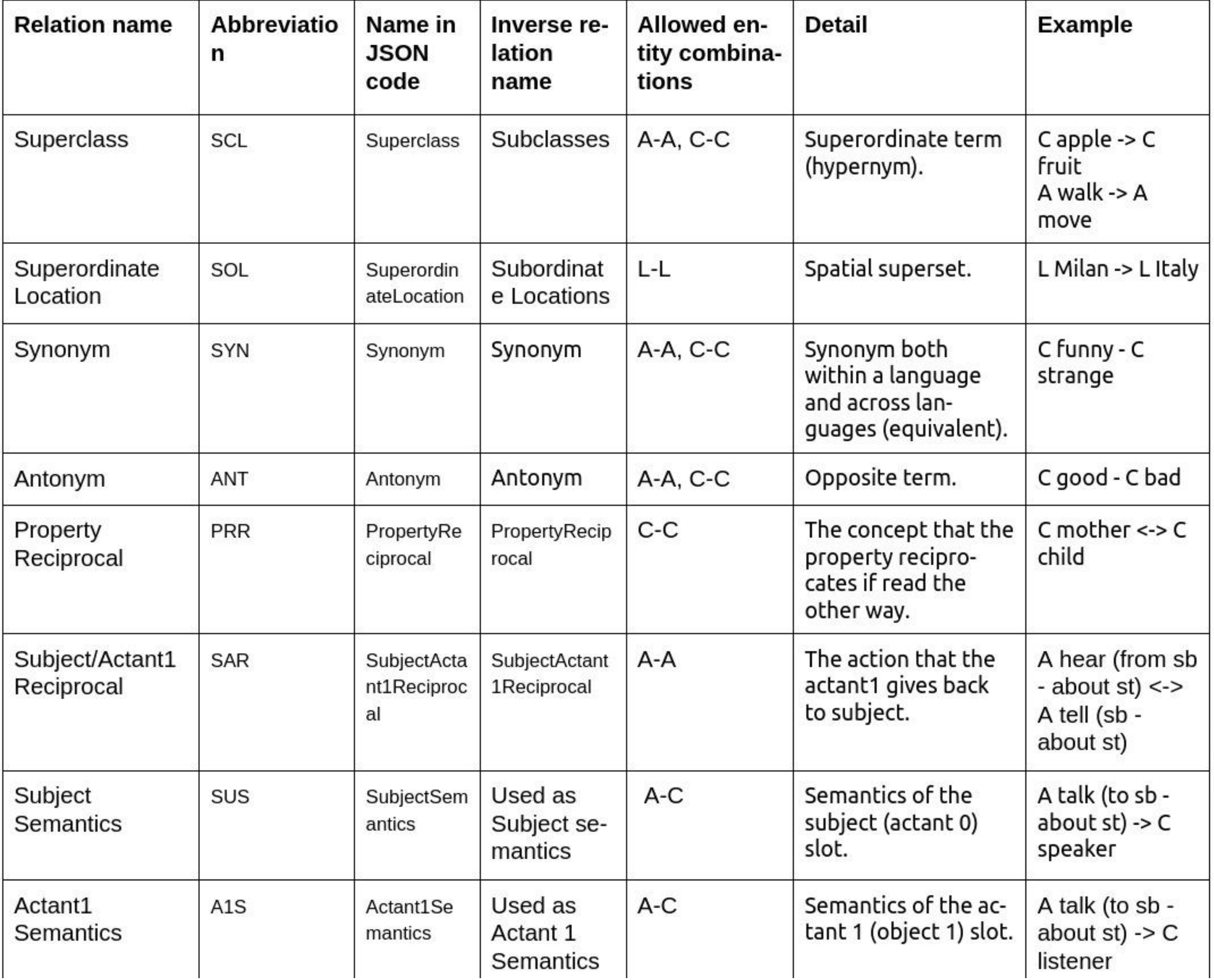
https://muni.cz/go/castemo

https://inkvisitor.net

dissinet.cz

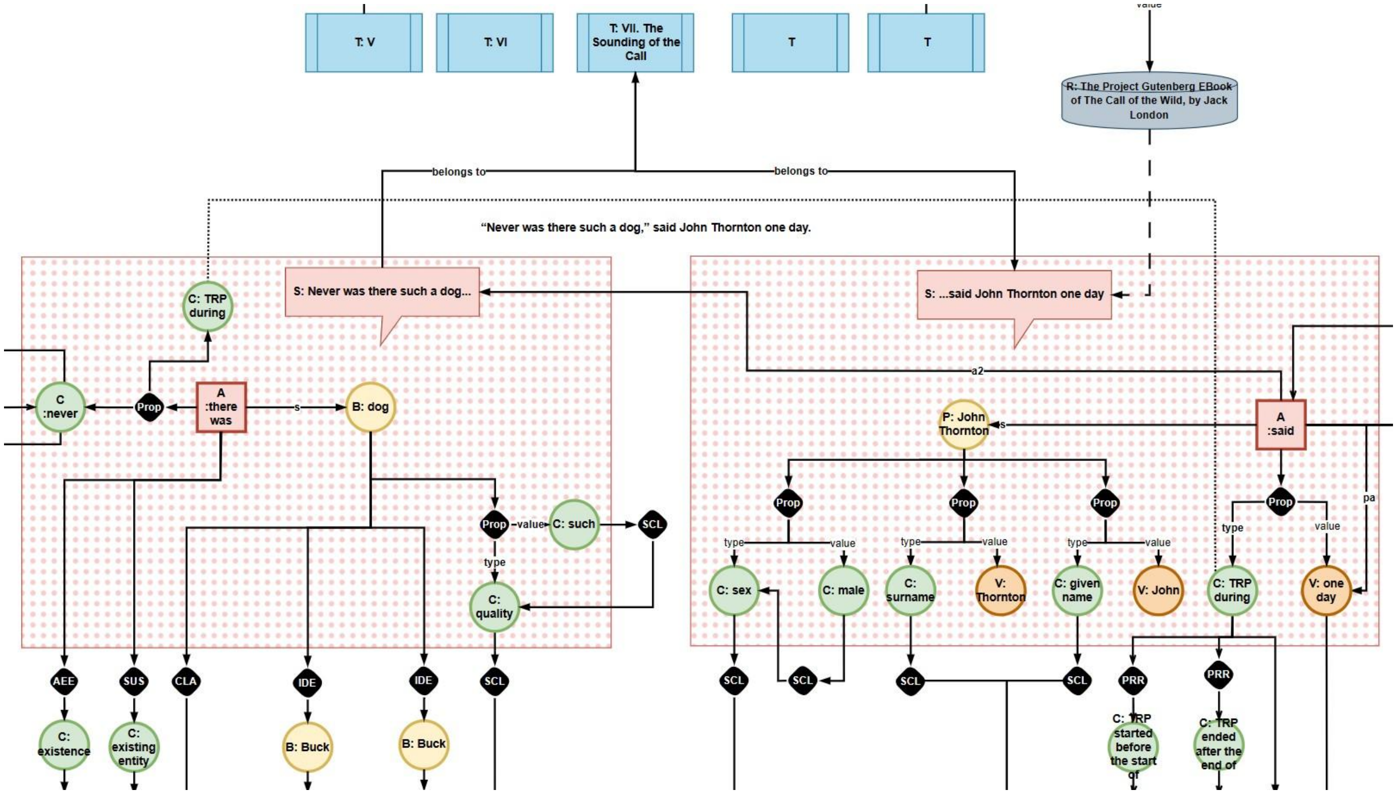
Bibliography

- Czech word sketches with dependency relations: https://www.sketchengine.eu/wp-content/uploads/Czech_word_sketch_2009.pdf
- Word Sketches for semantic relations: https://www.sketchengine.eu/wp-content/uploads/2016-Pattern-based-Word-Sketches-for-the-Extraction-of-Semantic-Relations.pdf



Relation name	Abbreviation	Name in JSON code	Inverse relation name	Allowed entity combinations	Detail	Example
Superclass	SCL	Superclass	Subclasses	A-A, C-C	Superordinate term (hypernym).	C apple -> C fruit A walk -> A move
Superordinate Location	SOL	SuperordinateLocation	Subordinate Locations	L-L	Spatial superset.	L Milan -> L Italy
Synonym	SYN	Synonym	Synonym	A-A, C-C	Synonym both within a language and across languages (equivalent).	C funny - C strange
Antonym	ANT	Antonym	Antonym	A-A, C-C	Opposite term.	C good - C bad
Property Reciprocal	PRR	PropertyReciprocal	PropertyReciprocal	C-C	The concept that the property reciprocates if read the other way.	C mother <-> C child
Subject/Actant1 Reciprocal	SAR	SubjectActant1Reciprocal	SubjectActant1Reciprocal	A-A	The action that the actant1 gives back to subject.	A hear (from sb - about st) <-> A tell (sb - about st)
Subject Semantics	SUS	SubjectSemantics	Used as Subject semantics	A-C	Semantics of the subject (actant 0) slot.	A talk (to sb - about st) -> C speaker
Actant1 Semantics	A1S	Actant1Semantics	Used as Actant 1 Semantics	A-C	Semantics of the actant 1 (object 1) slot.	A talk (to sb - about st) -> C listener

CASTEMO Data Model



CASTEMO

- Coding (annotating) **transcribed medieval inquisition** registers using the recently developed **Computer-Assisted Semantic Text Modelling (CASTEMO)**.
- Inspired by well-known ideas on meaning representation: RDF, OWL, Semantic Web, Quantitative Narrative Analysis
- Human-controlled, computer assisted, modeling the source closely: lexically (original language), syntactically, semantically, and contextually (part of document, context, textual order).
- Semantic units called statements that allow to model virtually everything in the source text
  - semantics, syntax, discursive elements, analytical layers (epistemic level, certainty, modality), conflicts, ambiguous evidence