TREEBANKING

Anastasia Mellano
Email: anastasia.mellano@gmail.com

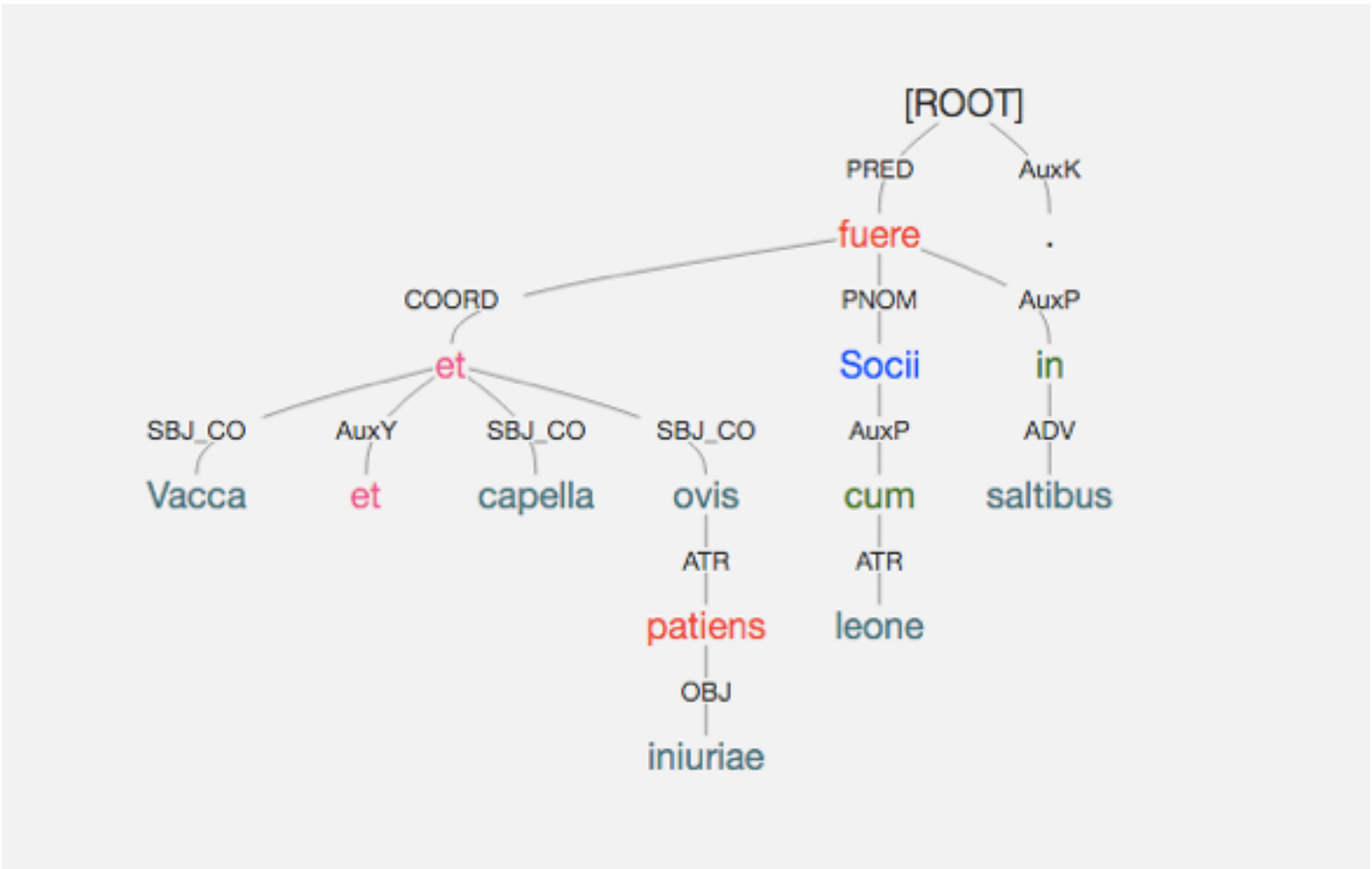
Introduzione

TREEBANKING: l'annotazione linguistica di un testo a livello morfologico, sintattico e semantico.

TREEBANK: il testo annotato, ossia fornito di uno specifico set di metadati.

Treebank

Phaedrus, *Fabulae* 1, 5:
Vacca et capella et patiens ovis iniuriae
socii fuere cum leone in saltibus.



Strumenti

- Prague Dependency Grammar, Universal Dependencies…

- Dizionari e grammatiche

- Piattaforme informatiche (parser) basate su uno specifico linguaggio di markup descrittivo (xml): Arethousa, Ugaritt (Perseids), Index Thomisticus, PROIEL, Penn Treebank…

- Linguaggio di programmazione: XQuery, XPath, Oxygen, Python, Java…

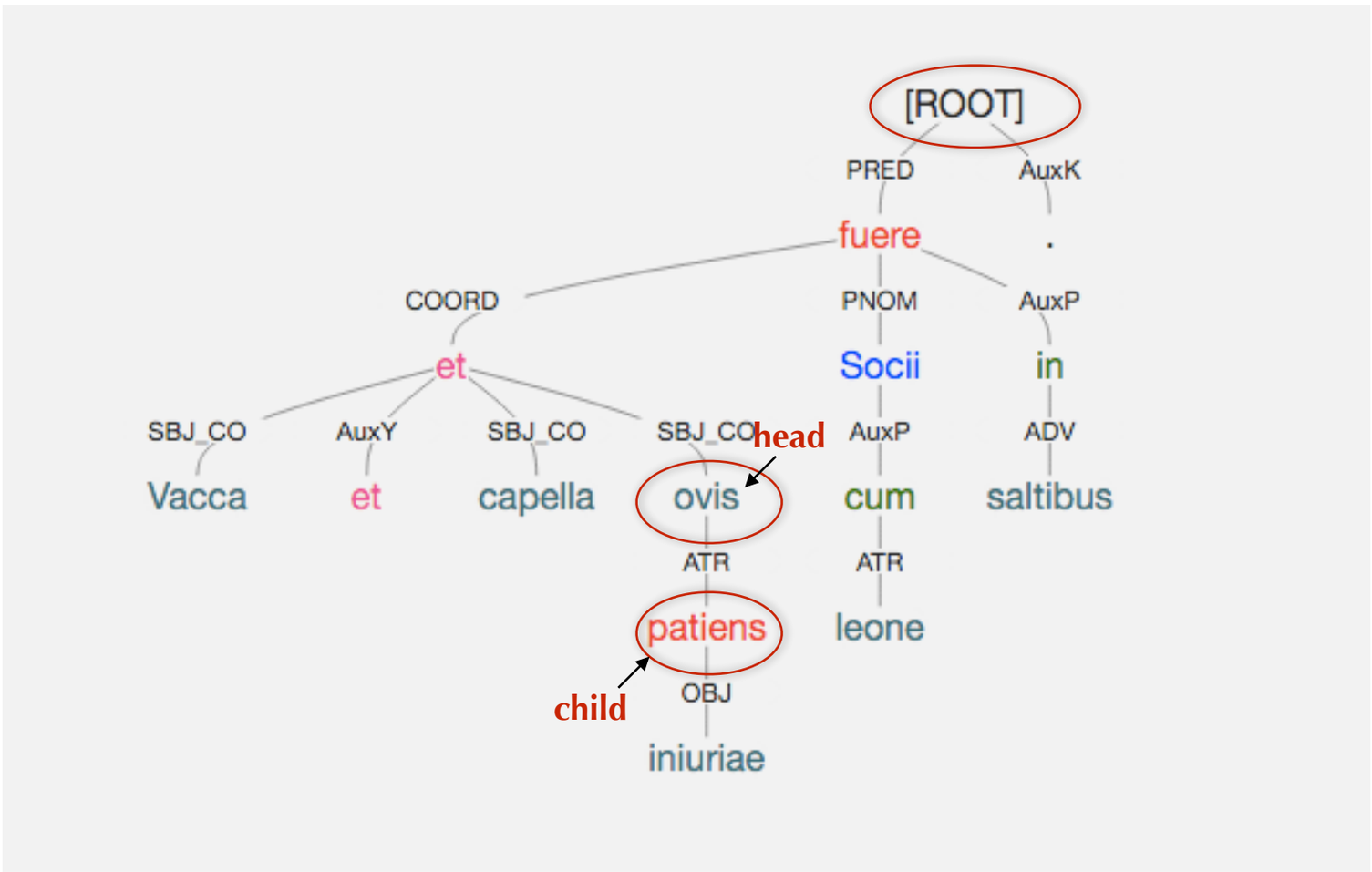
Grammatica di dipendenza

- Radice (root): la radice da cui si sviluppa l'albero. Ad essa si lega sempre il verbo principale, da cui dipendono in maniera diretta o indiretta tutti gli altri elementi del periodo

- Testa (head): elemento testuale che nel periodo regge un'altra unità testuale.

- Figlio (child): elemento testuale che nel periodo dipende da un'altra unità testuale.

Phaedrus, *Fabulae* 1, 5:
Vacca et capella et patiens ovis iniuriae socii fuere cum leone in saltibus.



Annotazione
(Arethousa)

La costruzione di un treebank

- Introduzione del testo all'interno del parser, che lo suddivide automaticamente in singoli periodi.

- Annotazione del testo secondo i diversi livelli di analisi prevista (morfologica, sintattica, semantica).

Annotazione

Avviene attraverso l'impiego di una serie di etichette distintive (tagset), che vengono associate agli elementi testuali secondo le regole della dependency grammar. A livello morfologico l'annotazione è semiautomatica e si distinguono undici categorie:

- verb (mood, tense, person, number, voice)
- noun (casus, gender, number)
- pronoun (casus, gender, number)
- adjective (casus, gender, number, degree)
- adverb (degree)
- conjunction
- preposition
- numeral (casus, gender, number)
- exclamation
- punctuation
- irregular

A livello sintattico l'annotazione è manuale e si individuano dodici macro-categorie:

- PRED (predicate): verbo principale della proposizione. Di norma è unico e si lega alla radice dell'albero.
- SBJ (subject): soggetto del verbo. Si può legare non solo al PRED, ma anche agli altri verbi del periodo.
- OBJ (object): argomento del verbo. Non è rappresentato soltanto dal complemento oggetto, ma da tutti gli argomenti (valenze) del verbo a cui si riferisce.
- ATR (attribute): attributo. Di norma si tratta di un elemento nominale, che per lo più si lega a teste nominali (sostantivi, pronomi e aggettivi). Fornisce informazioni aggiuntive e non necessarie.
- ADV (adverb): elemento avverbiale. Può essere di natura verbale o nominale e apporta informazioni aggiuntive all'interno del periodo. Si lega a verbi, sostantivi e avverbi. Non deve essere confuso con OBJ, perché la sua presenza è sempre opzionale.
- ATV/ATVv (verbal attribute): elemento nominale (sostantivo, aggettivo) che modifica in senso avverbiale la funzione del verbo a cui si lega.
- PNOM (predicate nominals): è sempre collegato a una testa verbale.
- OCOMP (object complement): è sempre collegato a una testa verbale.
- ExD: definisce gli incisi e il complemento di vocazione.

Bridge Structures

- COORD (coordinator): qualifica le congiunzioni coordinanti e unisce gli elementi che sono tra loro coordinati.

- APOS (apposition): definisce l'apposizione di un elemento. Di norma è un sostantivo, ma può anche essere rappresentata da un'intera proposizione.

- Auxiliary Elements: elementi sintatticamente poco consistenti, che nel periodo apportano informazioni aggiuntive oppure sono associati agli elementi portanti. Si suddividono in categorie minori: AuxP, AuxC, AuxV, AuxK, AuxX, AuxG, AuxY, AuxZ (preposizioni, congiunzioni subordinanti, verbi ausiliari, punteggiatura forte e debole, avverbi e particelle enfatiche).

Esercizio

Phaedrus, *Fabulae* 1, 5:
Vacca et capella et patiens ovis iniuriae
socii fuere cum leone in saltibus.

Obiettivi e prospettive di ricerca

- Didattica: rafforzamento delle conoscenze linguistiche attraverso una rappresentazione grafica e immediata del periodo e dei suoi rapporti sintattici.

- Ricerca: trattamento computazionale dei linguaggi naturali con riflessi sia a livello teorico (analisi delle principali strutture morfo-sintattiche, della loro ricorrenza e della loro evoluzione diacronica; comparazione tra differenti sistemi linguistici; studio delle interazioni psicolinguistiche) sia a livello pratico (creazione di macchine per la traduzione automatica, raccolte di dati e metadati….)

Bibliografia e sitografia

Progetti:

- Perseids: http://perseids.org
- The Penn Treebank: https://www.cis.upenn.edu/~treebank/

Istruzioni morfo-sintattiche:

- *A Latin Dictionary Founded on Andrews' Edition of Freund's Latin Dictionary. Revised, Enlarged, and in great part Rewritten by Charlton T. Lewis and Charles Short,* Oxford, Clarendon Press, 1955 (http://www.perseus.tufts.edu/hopper/resolveform?redirect=true&lang=Latin).
- David Bamman, Marco Passarotti, Gregory Crane, Savina Reynaud, *Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3)*, 2007, http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf
- James B. Greenough, George L. Kittredge, Albert A. Howard, Benjamin D'Ooge, *Allen and Greenough's New Latin Grammar for Schools and Colleges. Founded on Comparative Grammar*, New York, New Rochelle,1983 (http://www.documentacatholicaomnia.eu/03d/1903-1903,_Allen_and_Greenough,_New_Latin_Grammar,_EN.pdf e http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true).
- Ann Taylor, Mitchell Marcus, Beatrice Santorini, *The Penn Treebank: An Overview,* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.8216&rep=rep1&type=pdf

Repositories:

- Latin Treebanked Data: https://perseusdl.github.io/treebank_data/
- English Treebanked Data: https://www.ldc.upenn.edu