Linguistic annotation

Pesquisa de Informação em corpora, Lecture 8,
14 November 2012
Michel Généreux

I am grateful to Iris Hendrickx for letting me use and adapt her course material.

Overview

- Why?
- Tokenization
- **Syntactic linguistic annotation**
- **Semantic linguistic annotation**
- Stylistic annotation

Why Linguistic annotation?

- Make linguistic information explicitly visible
- **Helps with linguistic research questions**

  e.g. Adding info on lemmas makes it possible to search a corpus for a particular lemma

- Train automatic natural language processing (NLP) tools on the annotated corpus

Linguistic annotation

- Manual
- Semi-automatic
- Automatic approach
  - Symbolic: rule-based, lexicon-based
  - Statistical:

  Such automatic NLP tool needs to learn from manually annotated examples, but can then be used to label automatically large amounts of new text

Tokenization

- Recognition of word and sentence boundaries
- Basic step preceding linguistic annotation
- Usually done automatically
- Easy task:
  - Words are separated by spaces
  - Punctuation marks can be separated
  - Sentences start with Capital letter
  - Sentences end with . ? !

Tokenization Task

O Défice orçamental será de 2,8% da Produção final do País (PIB) em 2013, quando a economia crescer 1,7%, de acordo com o Programa de Estabilidade e Crescimento, PEC 2010–2013.

tokenized version:

<s id=1>O Défice orçamental será de 2,8% da Produção final do País (PIB) em 2013 , quando a economia crescer 1,7% , de acordo com o Programa de Estabilidade e Crescimento , PEC 2010 – 2013 .</s>

Hard cases: How to do:

- Numerical data *16.1°, 2007/2008, 16.345,95*
- To parse hyphens (-) *fim-de-semana, vestir-se*
- Portuguese contractions
  - Prepositions: *pela, numa, no*
  - Pronouns: *lho*
  - Verbs: *dar-lhos-ia*
- Ambiguity

  nos --> pronoun or 'em + os'

  deste --> verb 'dar' or 'de + este'

Hard cases(2)

Multiple words that form a unit:

- Multi word expressions

  depois de (after)

  botar um ponto final em (put an end to)

  to kick the bucket (to die)

- Compound words

  email address email-address

  fato de banho (swimsuit)

Overview

- Tokenization
- **Syntactic linguistic annotation**
  - **Part-of-speech**
  - **Parsing**
  - **Chunking**
  - **Lemmatization**
- Semantic linguistic annotation
- Stylistic annotation

Part-of-speech tagging

- Assign each word in the corpus its grammatical label in its context.
- Basic and most commonly used type of linguistic annotation -> precedes other steps of more complex annotation

POS-tagger: **automatically** predicts POS labels for words in a text

Nowadays, pretty good: 97% accurate for English

Example POS

'/PNT Haverá/V outro/ADJ encontro/CN esta/DEM noite/CN para/PREP tentar/INF finalizar/INF um/UM documento/CN conjunto/ADJ '/PNT ./PNT

CINTIL Corpus Tagset:

http://cintil.ul.pt/cintilwhatsin.html#pos

Excerpt from CINTIL tagset

Tag	Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz, ...
ADV	Adverbs	hoje, já, sim, felizmente, ...
CARD	Cardinals	zero, dez, cem, mil, ...
CJ	Conjunctions	e, ou, tal como, ...
CL	Clitics	o, lhe, se, ...
CN	Common Nouns	computador, cidade, ideia, ...
DA	Definite Articles	o, os, ...
DEM	Demonstratives	este, esses, aquele, ...
DFR	Denominators of Fractions	meio, terço, décimo, %, ...
DGTR	Roman Numerals	VI, LX, MMIII, MCMXCIX, ...
DGT	Arabic Numerals	0, 1, 42, 12345, 67890, ...
DM	Discourse Marker	olá, ...
EADR	Electronic Addresses	http://www.di.fc.ul.pt, ...
EOE	End of Enumeration	etc
EXC	Exclamation	ah, ei, ...
GER	Gerunds	sendo, afirmando, vivendo, ...
GERAUX	Gerund "ter"/"haver" in compound tenses	tendo, havendo
IA	Indefinite Articles	uns, umas, ...
IND	Indefinites	tudo, alguém, ninguém, ...
INF	Infinitive	ser, afirmar, viver, ...

Part-of-speech

Can range from just +/- 10 broad categories to fine-grained information about the verb inflection, gender or number details:

- metros/CN#mp: common noun, masculin, plural
- criança/CN#fs: common noun, feminin, singular
- abateu/V#ppi-3s: verb, past tense, 3rd person singular

Excerpt from CINTIL tagset

Tag	Description
Tags for nominal categories	
m	Masculine
f	Feminine
s	Singular
p	Plural
dim	Diminutive
sup	Superlative
comp	Comparative
Tags for verbs	
1	First Person
2	Second Person
3	Third Person
pi	Presente do Indicativo

POS-tagger

Many systems available, for many languages

Online demo:

- http://ucrel.lancs.ac.uk/claws/trial.html

  (Lancaster university)

- http://www.lsi.upc.edu/~nlp/SVMTool/demo.php
  (Universitat Politècnica de Catalunya

- http://lxcenter.di.fc.ul.pt/

LX-CENTER

LX suite

Developed at the University of Lisbon, Dept. of Informatics, by the NLX-Natural Language and Speech Group.

see an example | features | tagset | +services

Enter text in Portuguese, separating paragraphs with an empty line:

Haverá outro encontro esta noite para tentar finalizar um documento conjunto.

Annotate Clear

```
<p><s> Haverá/HAVER/V#fi-3s outro/OUTRO/ADJ#ms
encontro/ENCONTRO/CN#ms esta/DEM#fs noite/NOITE/CN#fs
para/PREP tentar/TENTAR/V#inf-nInf finalizar/FINALIZAR
/V#INF-nInf um/UM#ms documento/DOCUMENTO/CN#ms
conjunto/CONJUNTO/ADJ#ms .*//PNT </s></p>
```

Lemma

Lemma: canonical form of a word

For example:

- Verbs: infinitive *DAR* for forms *dava, dessem, dêem*
- Portuguese nominals: *masculin singular form*

e.g. gatas/GATO/CN#fp

Ambiguity problem:

- desse: DAR or DE+ESSE ?
- canto: cantar or canto ?

Lemma Lexicon

- ginja 3 ginginha ginja ginjas
- girafa 2 girafa girafas
- girândola 2 girândola girândolas
- girar 54 gira girada giradas girado girados girai girais giram giramos girámos girando girar girara girará giraram giráramos girarão giraras girarás gir ... etc.

CINTIL Example

A	A	DA#fs
programação	PROGRAMAÇÃO	CN#fs
e	e	CJ
a	a	DA#fs
condução	CONDUÇÃO	CN#fs
de_	de_	PREP
a	a	DA#fs
rega	REGA	CN#fs
desempenham	DESEMPENHAR	V#pi-3p
um	um	UM#ms
papel	PAPEL	CN#ms
importantissimo	IMPORTANTE	ADJ#ms-sup
em	em	PREP
condições	CONDIÇÃO	CN#fp
de	de	PREP
seca	SECA	CN#fs
.	.	PNT

Parsing

Analyse a sentence into its syntactic constituents

Annotation depends on the grammar type:

- Phrase-structure grammar
- Dependency grammar
- Functional grammar
- Transformational grammar

different theories, different annotations

Treebanks

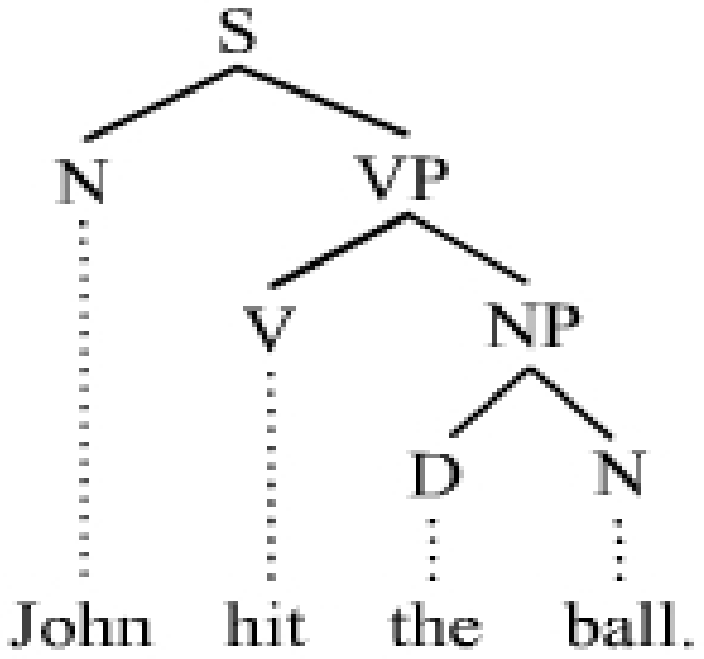
Corpus annotated fully with syntactic info is called **a treebank**

- Automatic Parsing is possible but to create a Treebank -> need a round of manual correction
- Some Treebanks are completely manual e.g. Lancaster-Leeds treebank

Parse example

John hit the ball.

(S (NP John)(VP(V hit)(NP(D the)(N ball))))



Constituency-based parse tree

Parsing: Demos

Stanford parser: English, dependencies and phrase structure http://nlp.stanford.edu:8080/parser/

- More info on parser: http://nlp.stanford.edu/software/lex-parser.shtml

Enju parser: English, HPSG grammar

http://www-tsujii.is.s.u-tokyo.ac.jp/enju/demo.html

Parsing for Portuguese

- Floresta Treebank with VISL tools

  http://beta.visl.sdu.dk/visl/pt/index.php

- LX Parser Based on Stanford parser, trained on 5000 manually annotated sententes

  http://lxcenter.di.fc.ul.pt/services/online_parser/features.html

Text Chunking

Dividing sentences into non-overlapping phrases = shallow parsing

Intermediate step: POS -> chunking-> parsing

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Demo: http://www.clips.ua.ac.be/cgi-bin/webdemo/MBSP-instant-webdemo.cgi

Overview

- Tokenization
- Syntactic linguistic annotation
- **Semantic linguistic annotation**
  - **Word sense**
  - **Semantic roles**
  - **Discourse:**
    - **Co-reference**
    - **Pragmatic information**
- Stylistic annotation

Semantic analysis

- How are the parts of a sentence or text related semantically?
- At multiple levels:
- Words: word sense disambiguation
- Sentences: semantic parsing, semantic roles
- Discourse: text structure representation, coreference, dialogue acts

Word Senses

- Semantic annotation at word level
- Types of problems:

Homonymy = group of words that share the same spelling and the same pronunciation but have different meanings

head (leader) vs. head (body part)

Polysemy = words with the same spelling and distinct but related meanings

blood bank vs. national bank

Vagueness:

nurse (male) or nurse (female)

Word Senses (2)

- Many words have more than one sense (meaning)
- The correct sense has to be decided on the basis of contextual and lexical information = disambiguation
- --> Lexical-semantic resource needed!

Use Dictionary?

action, noun *(Longman Dictionary)*

- Movement using force or power for some purpose
- Something done, deed
- The way in which a body moves
- The way in which a part of the body or a machine works
- The moving parts of a machine or instrument
- Effect
- A charge or a matter for consideration by a court of law
- + 6 more …

Dictionary: Problematic!

- Dictionary word definitions: circular! in terms of other words
- Who decides? Longman?
- Selection of senses seems to be arbitrary to large extent
- Some separated senses are more similar than others

Lexical Semantic Representations

- Constructive resources:

WordNet – EuroWordNet

- Online available: http://wordnet.princeton.edu

Meaning of words defined as:

- Synsets: a set of one or more synonyms
- Hypernyms – hyponyms
  - a hyponym shares a type-of relationship with its hypernym

WordNet – Example

- **bass** -- (the lowest part of the musical range)
- => low pitch, low frequency -- (a pitch that is perceived as below other pitches)
- => pitch -- (the property of sound that varies with variation in the frequency of vibration)
- => sound property -- (an attribute of sound)
- => property -- (a basic or essential attribute shared by all members of a class; "a study of the physical properties of atomic particles")
- => attribute -- (an abstraction belonging to or characteristic of an entity)
- => abstraction -- (a general concept formed by extracting common features from specific examples)

WordNet – *Twist*:

- # [11] turn of events
- # [10] an interpretation of a text or action
- # [04] device, gimmick,
- # [04] spin, twirl, twist, twisting, whirl (act of rotating rapidly)
- # [26] pull (a sharp strain on muscles)
- # [25] kink,
- # [25] bend, crook, twist, turn (a circular segment of a curve)
- # [04] social dancing

WordNet.PT

Online available:

http://www.clul.ul.pt/clg/wordnetpt/index.html

O banco: 3 senses

- car seat
- financial institute
- emergency room

Available resources

- To find corpora annotated with lexical semantic information:
- See: http://www.senseval.org/
  - Word sense multilingual annotation
  - semantic roles,
  - subcategorization information,
  - nouns compounds,
  - discourse etc.

Semantic annotation at sentence level

A sentence can be viewed as expressing or communicating about an event.

- **Thematic role**: meaning of noun in an event

The girl kissed the boy

- **agent** (causer of event): girl
- **theme** (beneficiary): boy

Ben gave the book to Cindy

- agent theme goal

Thematic roles: Problems

- People do not agree on defining one standard set of general roles

different theories, different annotations

- Different verbs restrict nouns in different ways

One can define:

- Generalized semantic roles
- Specialized semantic roles

Generalized Semantic Roles

Generalized roles that have certain characteristics. Arguments that have most of the properties are classified as such. For example:

- PROTO-AGENT
- PROTO-THEME

Propbank

Corpus annotated with Predicate-Argument structures (sem. roles)

- Both general sem. roles and verb-specific roles (e.g. *amount*)
- ARG0 = proto-agent
- ARG1 = proto-theme

[The oil price] has declined [10%].

ARG1 ARG2 (amount)

http://verbs.colorado.edu/propbank/framesets-english/

In Summary

Semantic annotation at sentence level

- People don't agree on semantic representations
- You saw one example of such a resource: Propbank

- Next: annotation at discourse/text level

Referring expressions

- Linguistic expressions like names and pronouns refer to referents

Referents can be :

- Entities
- Events

Example corefence chains

Kweli gave birth at the Bronx Zoo only after some meticulous matchmaking. Her calf, M'bura, was introduced to visitors last week. “It’s a zebra horse!” said Dagney Donaldson,7. Not exactly. Okapis are the only known relative of the giraffe, but with the silhouette of an antelope. They are notoriously skittish.

To breed such an antisocial animal in captivity requires a mix of patience, genetic know-how and romantic savoir-faire. So as M’bura, now a healthy 5-month-old, gallops around her shady habitat in the Bronx, zoo officials are taking a victory lap of their own. “There’s a lot of science in it, so you collect the data and use the data as a guide,” said Jim Breheny, director of the Bronx Zoo.

NYTimes, November 7, 2011

Example corefence chains (2)

Kweli gave birth at the Bronx Zoo only after some meticulous matchmaking. Her calf, M'bura, was introduced to visitors last week. "It's a zebra horse!" said Dagney Donaldson,7. Not exactly. Okapis are the only known relative of the giraffe, but with the silhouette of an antelope. They are notoriously skittish.

To breed such an antisocial animal in captivity requires a mix of patience, genetic know-how and romantic savoir-faire. So as M'bura, now a healthy 5-month-old, gallops around her shady habitat in the Bronx, zoo officials are taking a victory lap of their own. "There's a lot of science in it, so you collect the data and use the data as a guide," said Jim Breheny, director of the Bronx Zoo.

Discourse model

Webber (1978): people form mental model of ongoing discourse.

Simple mapping approaches for referents:

- FIFO
- LIFO
- Same semantic roles

Types of referring expressions

- **Indefinite NPs**: introduce new entity in discourse model

A colleague of mine has a mini.

- **Definite NPs:** refer to *known* entity
- *known:* mentioned before or strongly identifiable

Please close the door!

The moon is bright tonight.

Types of referring expressions (2)

- **Pronouns**: refer to previously mentioned entity. Pronouns can only refer to active references, salient entities

Kweli gave birth at the Bronx Zoo only after some meticulous matchmaking. Her calf, M'bura, was introduced to visitors last week. “It’s *a zebra horse*!”

Which is the most likely referent of 'it' ?

Annotation example

A 21-year old drunk decided monday evening to sleep out his drunkenness on the highway . The police found the man sleeping on his steering wheel.

<COREF ID="3">A 21-year old drunk</COREF> decided monday evening to sleep out <COREF ID="5" TYPE="IDENT" REF="3">his</COREF> drunkenness on the highway .

<COREF ID="6" >The police</COREF> found<COREF ID="7" TYPE="IDENT" REF="3">the man</COREF> sleeping on <COREF ID="8" TYPE="IDENT" REF="3"> his </COREF>steering wheel.

Online demo

- Online demo of automatic coreference system :

http://lingpipe-demos.com:8080/lingpipe-demos/coref_en_news_muc6/textInput.html

Discourse annotations

Corpus annotation: **Pragmatic** information

Speech acts

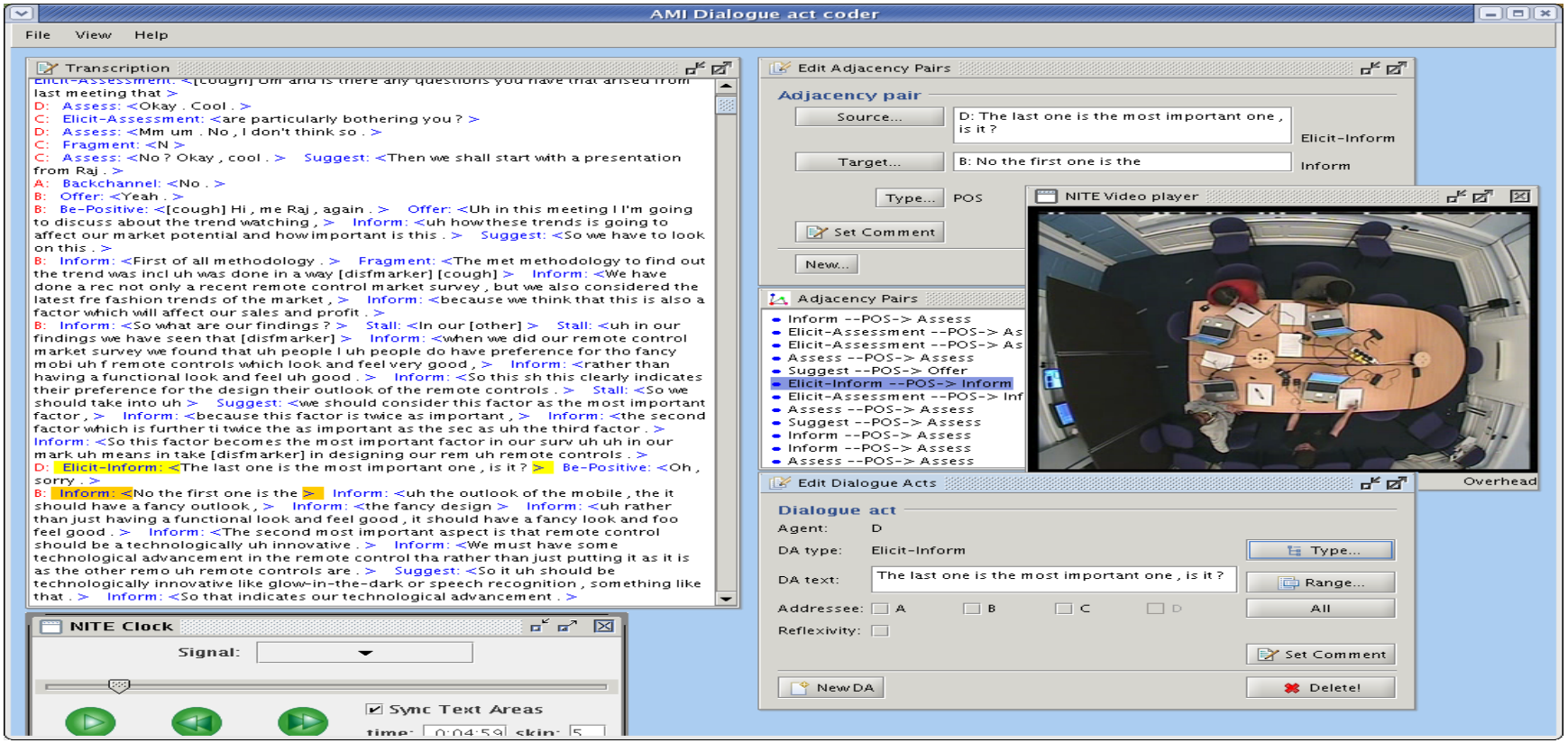
- apologies e.g. *sorry, excuse me*
- greetings e.g. *hello*
- hedges e.g. *kind of, sort of thing*
- politeness e.g. *please*
- responses e.g. *really?, that's right*

Speech Act Annotation

Example of speech act annotation in multi-modal data:

- *http://groups.inf.ed.ac.uk/nxt/Butterfly/Converted/AMI-dacoder.png*

Speech Act Annotation (2)



Stylistic Annotation

Annotate corpus with stylistic information

- Author style attributes

Cards: Cartas Desconhecidas corpus

http://alfclul.clul.ul.pt/cards-fly/index.php

- Type of speech (direct, indirect, etc.)

Lancaster Speech, Thought and Writing Presentation corpus

http://www.lancs.ac.uk/fass/projects/stwp/default.htm

Lancaster Speech, Thought and Writing Presentation corpus

<sptag cat=NRS who=B next=FIS whonext=B s=0.21 w=3>

I asked him — Narrative Report of Speech

</sptag>

<sptag cat=FIS who=B next=DS whonext=B s=0.43 w=6>

what Franco was doing down here. — Free Indirect Speech

</sptag>

<sptag cat=DS who=L next=NRS whonext=L s=0.64 w=7>

'He is opening the new Almeria airport,' — Direct Speech

</sptag>

<sptag cat=NRS who=L next=N s=0.36 w=4>

he said with pride.

</sptag>

Named Entity Recognition

Locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Jim bought 300 shares of Acme Corp. In 2006.

<ENAMEX TYPE="PERSON">Jim</ENAMEX>bought<NUMEX TYPE="QUANTITY">300</NUMEX>shares of<ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>.

IOB annotation for NPs

A	DA	B-NP	
caçadeira		CN	I-NP
,	PNT	I-NP	
em	PREP	I-NP	
princípio		CN	I-NP
o	DA	I-NP	
instrumento		CN	I-NP
do	PREP+DA	I-NP	
crime	CN	I-NP	
,	PNT	I-NP	
não	ADV	O	
está	V	O	
perto	ADV	O	
de	PREP	O	
qualquer		QNT	B-NP
dos	PREP+DA	I-NP	
corpos	CN	I-NP	
.	PNT	O	

Cards XML annotation

```
<seg type="verticalText">
  que deem o <abbr>dro</abbr> e se não as podes deser man
   dame diser e pera te servir faco vir
   <gap reason="damage" extent="1 word"/>
   <seg type="formulaicText" n="peroration">
        Deus te   <abbr>gde</abbr>
   </seg>
</seg>
<closer>
            <dateline>Agosto e ja dto. 29 de 1724</dateline>
            <salute>teu tio</salute>
            <signed>
                o <abbr>Pe</abbr>Jorge<abbr>frz</abbr>
             </signed>
</closer>
```

OpinionFinder XML annotation

<MPQASRC>Source of a speech event or private state </MPQASRC>
<MPQASD> Direct Subjective Expression and Speech Event</MPQASD>
<MPQAPOL>Polarity classifier</MPQAPOL>
<MPQASENT>Sentence classifier</MPQASENT>

Howard Moore, sent a copy of his book, The Universal Kinship, with a letter in which he said: "Most humorists have no anxiety except to glorify themselves and add substance to their pocket-books by making their readers laugh.

<MPQASENT autoclass1="subj" autoclass2="subj" diff="33.5">Howard Moore, sent a copy of his book, The Universal Kinship, with a letter in which <MPQASRC>he</MPQASRC> <MPQASD>said</MPQASD>: "Most humorists have no <MPQAPOL autoclass="negative"> anxiety </MPQAPOL> except to <MPQASD>glorify</MPQASD> themselves and add substance to their pocket-books by making their readers <MPQASD>laugh</MPQASD> . </MPQASENT>

Text Annotation

Gender, Age, Nationality, etc.

Moods

Sentiment (positive, negative)

Proxy Annotation: financial news <-> markets movements

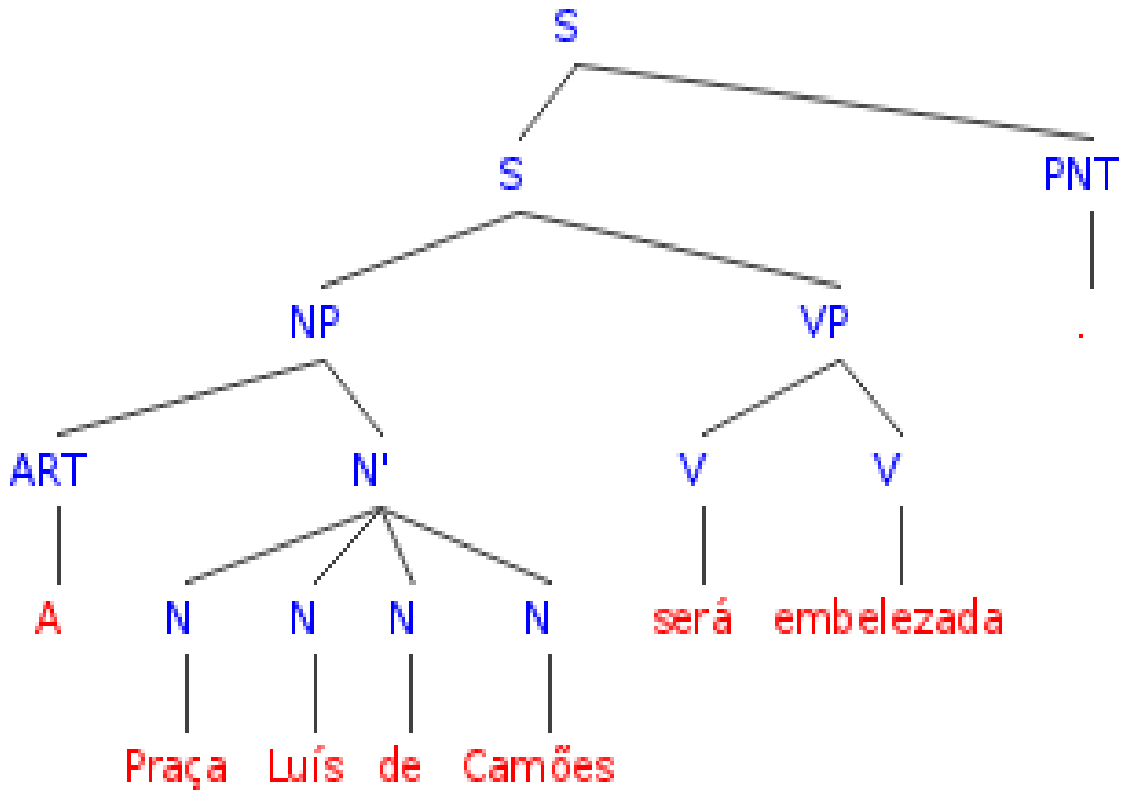
Useful links

Many resources, postaggers, chunkers, etc., listed here:

- *http://nlp.stanford.edu/links/statnlp.html*

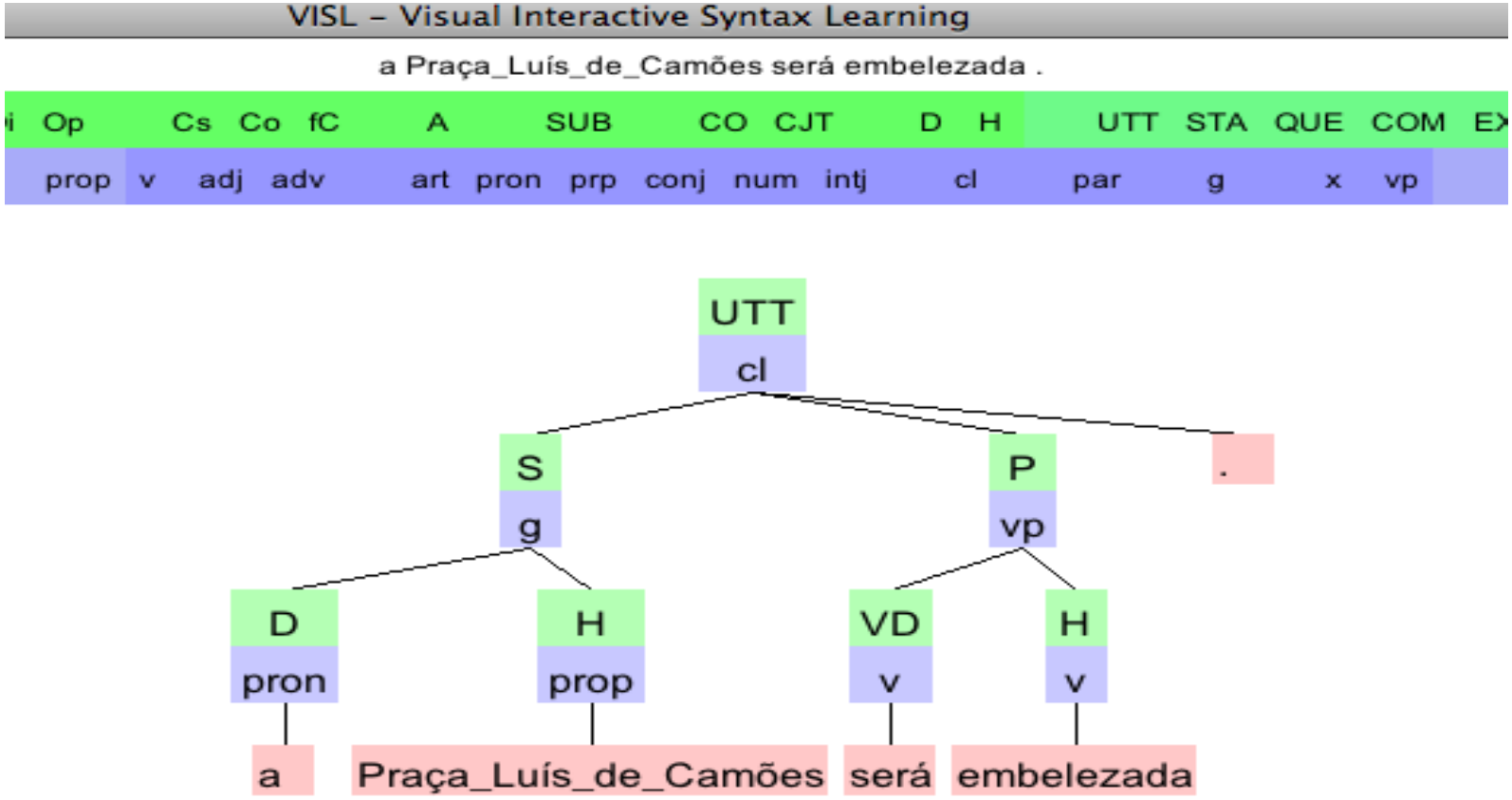
Parse tree LX

http://lxcenter.di.fc.ul.pt/services/en/LXServicesSuite.html



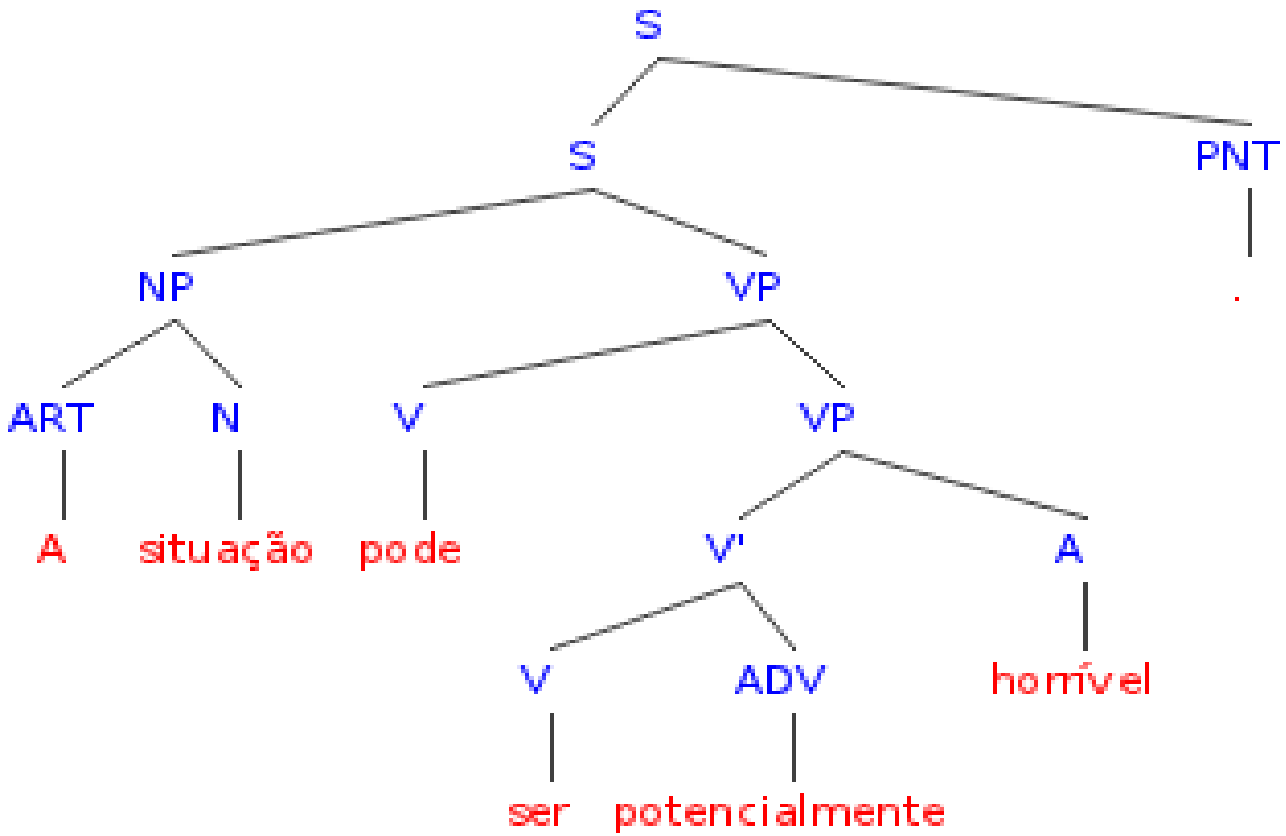
Parse Tree Floresta

http://beta.visl.sdu.dk/visl/pt/index.php



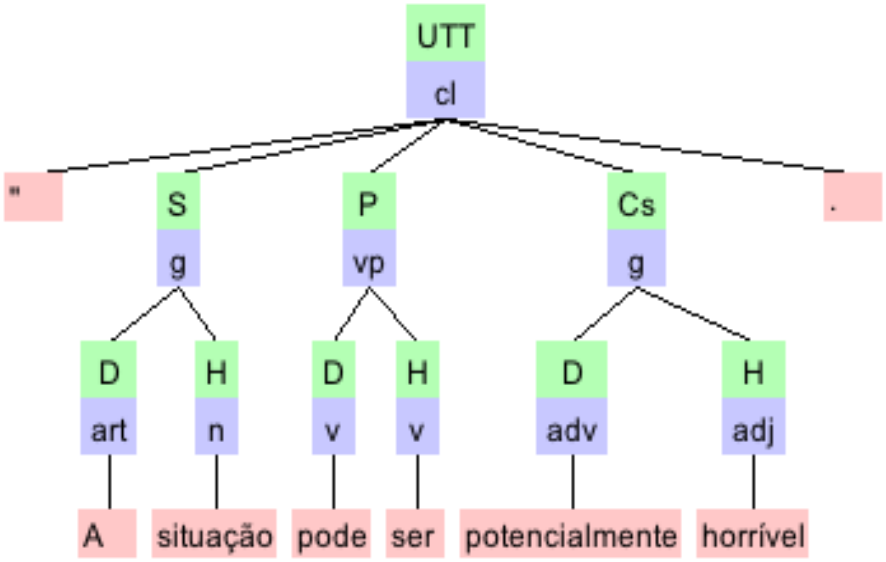
LX-parser

http://lxcenter.di.fc.ul.pt/services/en/LXServicesSuite.html



Parse Tree Floresta

http://beta.visl.sdu.dk/visl/pt/index.php



Stanford parser

http://nlp.stanford.edu:8080/parser/

Parse

```
(ROOT
  (S
    (NP (DT The) (NN girl))
    (VP (VBD ate)
      (NP (DT a) (NN pizza))
      (PP (IN with)
        (NP (NNS salami))))
    (. .)))
```

Stanford parser
http://nlp.stanford.edu:8080/parser/

Parse

```
(ROOT
  (S
    (NP (DT The) (NN girl))
    (VP (VBD ate)
      (NP (DT a) (NN pizza))
      (PP (IN with)
        (NP (DT a) (NN fork))))
    (. .)))
```

Wordnet-EN

http://wordnetweb.princeton.edu/perl/webwn

- **S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds)** ***"the dog barked all night"***
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *member holonym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
- S: (n) frump, **dog** (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"*
- S: (n) **dog** (informal term for a man) *"you lucky dog"*
- S: (n) cad, bounder, blackguard, **dog**, hound, heel (someone who is morally reprehensible) *"you dirty dog"*
- S: (n) frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) pawl, detent, click, **dog** (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) andiron, firedog, **dog**, dog-iron (metal supports for logs in a fireplace) *"the andirons were too hot to touch"*

Wordnet-PT

http://www.clul.ul.pt/clg/wordnetpt/index.html

cão (cat. nominal)

glosa: canídeo doméstico

"cão" é quase sinónimo de ão-ão (onomatopeia)

equivalência em Inglês

"cão" é hiperónimo (é supertipo) de

- pastor-alemão, lobo da Alsácia - cão robusto de pêlo curto em ton
- São Bernardo, cão de São Bernardo - cão corpulento de pêlo bran
- caniche - cão de cabeça fina e longa, pêlo frisado e orelhas compr
- Serra da Estrela, cão da Serra da Estrela - cão castanho de pêlo c
- Serra d'Aires, Serra de Aires, cão de Serra de Aires - cão de memb

Propbank

http://verbs.colorado.edu/propbank/framesets-english/

Roles:

Arg0: *lover* (vnrole: 31.2-1-experiencer)
Arg1: *loved* (vnrole: 31.2-1-cause)

Example: active transitive

Some institutional traders loved the wild ride .

Arg0: Some institutional traders
Rel: loved
Arg1: the wild ride

SentiWordnet
http://sentiwordnet.isti.cnr.it/

LingPipe
http://alias-i.com/lingpipe/

Demonstrators:

- Encoding
- Sentence splitting
- POS tagging
- Named Entity Recognition
- Coreference Resolution