Cristiano Chesi (c. chesi @ unisi. it)

ANNOTATED CORPUS CREATION (TREEBANK)

Goals

(1) Creating and exploring an annotated corpus in XML format
(2) Start using a semi-automatic tool for annotating a treebank

XMLTreeEditor

(3) Sample text annotated using **XML**

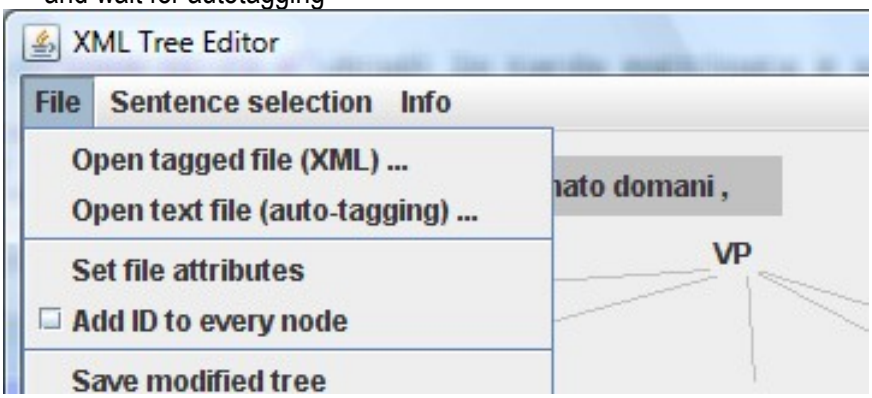
più difficile la situazione in Senato domani

```
<node cat="VP" id="2008-01-23.3" role="head">
        <node agree="f.s" cat="AP" id="1" lp="topic" role="arg.predobj">
                <word cat="ADV.streng" id="2" lemma="più">più</word>
                <word agree="f.s" cat="A.qualif" id="3" lemma="difficile">difficile</word>
        </node>
        <word agree="3.s" cat="V.ind.fut" id="4" lemma="essere" role="head" subcat="copula"/>
        <node agree="f.s" cat="NP" id="5" role="arg.subj">
                <word agree="f.s" cat="D.art.def" id="6" lemma="la">la</word>
                <word agree="f.s" cat="N.comm.count.inanim" id="7" lemma="situazione"
                role="head">situazione</word>
        </node>
        <node cat="NP" id="8" role="adj.loc">
                <word cat="P.loc" id="9" lemma="in">in</word>
                <word agree="m.s" cat="NE.org" id="10" lemma="senato"  role="head">senato</word>
        </node>
        <word cat="ADV.time" id="11" lemma="domani" role="head">domani</word>
        <word cat="END.comma" id="12" lemma=",">,</word>
</node>
```

(4) Create the corpus using XMLTreeEditor (few simple text file, ANSI encoding):

1. Dowonload Java Runtime Environment, JRE (http://www.java.com/it/download/index.jsp)
2. Download a simple text editor:
   Windows: "Programmers's Notepad" (http://www.pnotepad.org/)
   or "Notepad ++" (http://notepad-plus-plus.org/)
   Mac: TextWrangler (http://www.barebones.com/products/textwrangler/)
3. Downaload the XMLTreeEditor http://www.ciscl.unisi.it/master/materials.htm/xmltreeditor.zip
4. Download a tagged sample: http://www.ciscl.unisi.it/master/materials.htm/corpus-sample.zip
5. Use some text file (UTF-8) encoding and normalize the transcription (one sentence per line, check the orthography, named entities and meaning…)

6. Launch tool and select “Open text file (auto-tagging)” from the “File” menu; select the edited text, and wait for autotagging



7. Annotate errors, doubts, complex phrases.

(5) morphosyntactic phrases:

Nouns

e.g. “case” (houses): cat=“N.comm.count.inanim”, agree=“f.p”, role=“head” lemma=“casa”

Attribute	*Value (**default,** [optional])*	*Explaination*
Cat	1. N/N.pro[.cl] 2. [**comm**/prop] 3. [**count**/mass] 4. [anim/[per[.first/.last] /impers/reflex] /**inanim/**[city/gpe/org]]	noun/pronoun[clitic] common/proper contable/mass animate/[person[first/last name] impersonal/reflexive] /inanimate[city/geo-political entity/company]
Agree	1. [**m**/f/n] 2. [**s**/p/n]	masc/sing/neut gender sing/plur/null number
Role	**head**/arg/adj	head / selected argument / unselected adjunct
Sem	[alphanumeric index]	MultiWordnet id
Lemma	[any alphanumeric character]	dictionary uninflected form, if null its value is the token form

Cristiano Chesi (c. chesi @ unisi. it)

Verbs

e.g. “corre” ((he) runs): cat=“V.ind.pres”, agree=“s”, role=“head” lemma=“correre”)

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. V/V.aux/V.mod/V.asp 2. **ind**/subj/cond/part/imp/inf 3. pres/past/past+/fut/fut+/impf 4. [state/event[.atelic/.telic[.punct]]]	main/auxiliary/modal/aspectual verb indicative/subjunctive/conditional/ participe/imperative/inifitive mood present/past/remote past/future/ anterior future/imperfect aspectual classes (e.g. “cough” is an event, telic and punctual)
Subcat	transitive/intransitive/ditransitive/ unaccusative/copula/ causative/passive/psych/ control_subj/control_obj	Subcategorization classes
Agree	1. [1/2/**3**] 2. [m/f/**n**] 3. [**s**/p/n]	person gender number
Role	**head**/[adj]	head / unselected adjunct (e.g. auxiliaries, modals)

Adjectives

e.g. “forte” (strong): cat=“A.qualif”, agree=“f.s”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. A 2. deict/dem/excl/indef/interr/nation/ num[.ord/.card]/poss/qualif	adjective deictic/demonstrative/exclamative/ interrogative/geographical specification/numeral[ordinal/cardinal]/ possessive/qualificative
Subcat	super/dimin/compar	superlative/diminutive/comparative form
Agree	as for Nouns	
Role	as for Nouns	

Adverbs

e.g. *prima* (*before*): cat=“ADV.time”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. ADV 2. adfirm/advers/compar/doubt/ interr/limit/loc[.pro.cl]/manner/neg/ quant/reason/streng/ superl/temp	adverb adfirmirmative/adversative/comparative /doubitative/interrogative/limitative/ locative[.pro.cl]/manner/negative/ quantitative/reason/strength/ superlative/tempoparl
Role	[adj]	adjunct

Determiners

e.g. *il* gatto (*the* cat): cat=“D.art.def”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. D 2. art[.def/.indef]/demo/ quant[.univ/.exist/.comp/.distr/.neg]	determiner article[definite/indefinite]/demonstrative/ quantifier[universal/exististential/ comparative/distributive/negative]
Agree	Same as Nouns	
Role	[adj]	adjunct

Cristiano Chesi (c. chesi @ unisi. it)

Prepositions

e.g. “il libro *di* Gianni” (the book *of* G.): cat=“P.genitive”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. P 2. advers/benef/comitat/compar /dative/evident/genitive/goal /instr/loc/manner/malefact /material/matter/means/measure /partitive/path/reason/source/temp	adverb adversative/benefactive/comitative/ comparative/dative/evidential/genitive/ goal/instrument/locative/manner/ malefactive/material/matter/means/measure /partitive/path/reason/source/temporal
Role	[adj]	adjunct

Complementizers

e.g. “di” (to): cat=“C.decl”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. C 2. coord[.advers]/rel.pro/wh/ subord[.advers/.reason/.goal .conc/.cond/.decl/.fin/.loc/.temp]	complementaizer coordination[.adversative]/relative pronoun/wh-element/ subordinator[adversative/reason/goal concessive/conditional/declarative/ final/locative/temporal]
Role	[adj]	adjunct

Specials

e.g. “.” (dot, punctuation): cat=“END.period”

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. END/ABBR/INT/SPECIAL 2. period/comma/colon/scolon/quote	punctuation/abbreviations/interjections/ special characters (e.g. currency, percentage etc.)

Non terminal nodes

NPs, VPs and APs

Attribute	*Value (**default**, [optional])*	*Explaination*
Cat	1. NP/VP/AP/FRAG	nominal/verbal/modifier (both adjectival and adverbial) phrases/fragment
Role	adj	adjunct

(13) Dependencies:

use them to indicate relations among constituents:

- head — phase head
- arg(uments)
  - subj(ect) — nominative case-marked argument
  - obj(ect) — accusative case-marked argument
  - ind(irect)obj(ect) — third argument (e.g. dative)
  - predobj(ect) — object in copular constructions
- adj(uncts)
  - advers — adversative specification
  - adfirm — affirmative specification
  - benef — benefactive specification
  - cond — conditional specification
  - coord — coordination specification (second conjunct is marked adj.coord and it is dominated by the previous one)
  - comitat — comitative specification
  - compar — comparative specification
  - hangtopic — extra argument (topic) specification
  - measure — measure specification
  - evident — evidential specification
  - goal — goal specification
  - instr — instrument specification
  - loc — locative specification
  - malefact — malefactive specification
  - manner — manner specification
  - matter — matter specification
  - means — means specification
  - path — path specification
  - partitive — partitive specification
  - reason — reason specification
  - source — source specification
  - temp — temporal specification
  - rel — relative clause
    - restr — restrictive relative
    - adpos — adpositive relative

We decided to subcategorize prepositions according to the functional specification they introduce (the relation is not always 1-to-1). The following table summarizes the main subcategories briefly explaining them.

Prepositional subcategory	Examples	Brief Explanation [*Typically, it can be used to answers a question such as:*]
Genitive	*il presidente **della** repubblica* (arg.obj - i.e. a specification) [the president of the Republic] *la conferma **dei** socialisti* (arg.subj - i.e. subject/owner) [the confirmation of the Socialists] *le chiavi **di** casa* (adj.matter) [the keys of the house]	Usually used for animate complements, it introduces a specification or the subject or the owner of something [*of whom?*]
Matter	*risultati **delle** elezioni* (arg.obj) [the results of the elections] *rinunciare **alla** carica* (indobj) [to give up an office]	Usually used for inanimate complements, it introduces the matter or topic of something [*about/of what?*]
Dative	*essere ucciso **dai** carabinieri* (indobj - passive) [being killed by cops]	It introduces the indirect object
Loc	*vivo **a** Roma* [I live in Rome]	It introduces the place where the action occurs [*where did it happen?*]
Source	*uscire **di** casa* [to leave the house]	It introduces the origin of a movement [*from where does x move?*]
Path	*Vado **verso** la periferia* [I'm going towards the outskirts]	It introduces the direction of a movement [*towards what does x move?*]
Benef	*mese positivo **per** l'economia* [positive month for the economy]	It introduces the participant who benefits from the action [*for whom?*]
Malefact	*dare fuoco **al** pino* [to set fire to the pine tree]	It introduces an opponent, as well as a participant who is penalized by the action [*against whom/what?*]
Manner	*corro **da** solo* [I run by myself]	It introduces the manner in which a certain action takes place [*how?*]
Means	*vado **col** treno* [I move by train]	It introduces the mean of transportation [*by/with what?*]
Measure	*crescre **di** 3 metri* [to grow 3 meters]	It introduces a quantitative description of an action [*how much?*]
Temp	*dormo **da** giorni* [I slept for days] *pulisco **di** domenica* [I clean up on sunday]	It introduces a temporal characterization of an action [*When? How long? From when? Untill when?...*]
Comitat	*l'accordo **coi** centristi* [the deal with the centrists]	It introduces other people that share the role of the subject [*with whom?*]

Partitive	*uno **di** noi* [one of us]	It introduces the set which an object belongs to [*of what (set)?*]
Instrument	*lingua **dei** segni* [sign language - "a language that uses visually transmitted sign pattern"]	It introduces the object used to perform the action [*by using what?*]
Material	*la casa **di** legno* [the house made of wood]	It introduces the substance which an object is made of [*made of what?*]
Evident	***secondo** il Presidente* [according to the President]	It introduces someone perspective [*according to what/whom?*]
Compar	*più bello **di** me* [more beautiful than me]	It introduces the second term of a comparison [*compared to whom/what?*]
Reason	*accordo **per** il ballottaggio* [the deal for the ballots]	It introduces the cause of a certain action [*because of what?*]
Goal	*corsa **per** la vittoria* [running for victor]	It introduces the goal of an action [*why/for what?*]

Riferimenti

Cristiano Chesi, Gianluca Lebani, Margherita Pallottino (2008)
A Bilingual Treebank (ITA-LIS) suitable for Machine Translation: what Cartography and Minimalism teach us.
StIL Vol. 2
http://www.ciscl.unisi.it/doc/doc_pub/chesi-lebani-pallottino2008-A_Bilingual_Treebank_ITA-LIS_suitable_for_Machine_Translation.pdf