The Italian dependency annotated corpus developed for the CoNLL-X Shared Task ISST-CoNLL

Simonetta Montemagni – ILC-CNR
Maria Simi – Dipartimento di Informatica, Univ. Pisa

Abstract. This document illustrates the Italian dependency annotated corpus developed for the CoNLL-X Shared Task (henceforth referred to as ISST-CoNLL). In particular, it provides information on the background resource, the way the CoNLL Italian resource has been designed and developed, and finally documents the adopted annotation scheme.

The Italian dependency annotated corpus developed for the CoNLL-2007 Shared Task ISST-CoNLL

Abstract. This document illustrates the Italian dependency annotated corpus developed for the CoNLL-2007 Shared Task (henceforth referred to as ISST-CoNLL). In particular, it provides information on the background resource, the way the CoNLL Italian resource has been designed and developed, and finally describes the adopted annotation scheme.

1 The background resource

The Italian dependency annotated corpus developed for the CoNLL-2007 Shared Task was derived from the Italian Syntactic-Semantic Treebank (ISST), a multi-layered annotated corpus of Italian which represents one of the main outcomes of an Italian national project, SI-TAL, funded by the Italian Ministery of Science and Research (MURST) for the design and development of an integrated suite of tools and resources for Italian Natural Language Processing.

The project consortium included companies and computational linguistics sites in Italy which are active with different expertise in the computational linguistics field (ILC-CNR/CPR, Venezia University/CVR, ITC-IRST, “Tor Vergata” University/CERTIA and Synthema). ISST was developed between 1999 and 2001 and is being updated whenever needed.

It is worth mentioning here that among the expected uses for ISST there were also training (and/or tuning) of grammars and sense disambiguation systems and the evaluation of language technology systems. For more details see Montemagni et al. (2003a, 2003b) and the web page http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=874/vers=ing.

1.1 ISST overall architecture

ISST has a five-level structure covering orthographic, morpho-syntactic, syntactic and semantic levels of linguistic description. Syntactic annotation is distributed over two different levels: the constituent structure level and the dependency annotation level. The fifth level deals with lexico-semantic annotation, which is carried out in terms of sense tagging of lexical heads (nouns, verbs and adjectives) augmented with other types of semantic information: ItalWordNet is the reference lexical resource used for the sense tagging task. Both syntactic and lexico-semantic annotations refer to the morpho-syntactically annotated text, which in turn is linked to the orthographic file with the text and mark-up of macrotextual organisation (e.g. titles, subtitles, summary, body of article, paragraphs).

1.2 Syntactic annotation in ISST

Among the main features of ISST with respect to other treebanks there is the distributed approach to syntactic annotation. In this respect, ISST differs from most treebanks which adopt a unique syntactic representation layer. ISST also differs from multi-level treebanks like the Prague Dependency Treebank (PTD): whereas PTD annotation levels refer respectively to a) the surface dependency relations and b) the underlying sentence structure, ISST syntactic annotation levels are intended to provide orthogonal and independent views of the same surface syntax. None of the ISST syntactic annotation levels presupposes the other; on the other hand combined views of the complementary information contained in them can be provided, e.g. dependency information can be projected onto the constituent structure. The motivations underlying this “double-track” approach to syntactic representation range from language-specific ones (e.g. the syntactically free constituent order and the pro-drop property of Italian) to usage-oriented ones (ISST syntactic annotation levels are intended to be exploitable both in real applications and for research purposes, and to be compatible with different approaches to syntax, either adopted in theoretical or applicative frameworks).

1.3 Corpus composition

The ISST corpus consists of 305,547 word tokens reflecting contemporary language use.
It includes two different sections:

1) a “balanced” corpus, testifying general language usage, for a total of 215,606 tokens;
2) a specialised corpus, amounting to 89,941 tokens, with texts belonging to the financial domain.

The balanced corpus contains a selection of articles from different types of Italian texts, namely newspapers (*La Repubblica* and *Il Corriere della Sera*) and a number of different periodicals which were selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.) covering a 10 year time period (1985-1995). The financial corpus includes articles taken from *Il Sole-24Ore* which were published in 1994.

2 The CoNLL-2007 Shared Task Italian resource: ISST-CoNLL

ISST was used as the starting point to build the Italian annotated corpus developed for the CoNLL-2007 Shared Task. In particular, ISST-CoNLL was built on top of the morpho-syntactic and syntactic dependency annotation levels through a semi-automatic conversion process cooperatively carried out by ILC-CNR and the Dipartimento di Informatica of the University of Pisa.
Conversion was in charge of :

a) combining information coming from two different annotation levels
b) converting the ISST annotation scheme for dependency annotation into the CoNLL-2007 tabular format.

Conversion had to cope with the fact that in ISST dependency relations are expressed in terms of binary relations holding between two lexical heads belonging to major lexical classes only (i.e. non-auxiliary verbs, nouns, adjectives and adverbs): in fact, in ISST information about grammatical words (e.g. determiners, prepositions, auxiliaries) is encoded in terms of features associated with the participants to the relation. This implies that during the conversion process the dependency relations involving grammatical words had to be reconstructed from the ISST original annotation and the already existing dependency relations had to be revised accordingly.
This was done semi-automatically by means of several conversion scripts whose output has been manually revised with the help of a graphical annotation tool. Further scripts were run to validate the consistency of the final output. An XML intermediate format was produced in this process, with purpose of preserving original annotations that could not be accomodated in the CoNLL format.
For what concerns corpus composition, ISST-CoNLL is a subset of the balanced ISST corpus of 79654 word tokens (of which 65016 were non punctuation tokens) for a total 4162 sentences, corresponding to the *Corriere della Sera* and *periodicals* partitions.

ISST-CoNLL is copyrighted material which can be used for research purposes only and which cannot be distributed in any original or modified form (see the licence agreement form).

3 ISST-CoNLL annotation

3.1 Data format

Annotated data adheres to the following general rules:

- data files contain sentences separated by a blank line;
- a sentence consists of one or more tokens, each one starting on a new line;
- a token consists of ten fields (namely ID, FORM, LEMMA, CPOSTAG, POSTAG, FEATS, HEAD, DEPREL, PHEAD and PDEPREL) separated by a single tab character;
- all data files will contain for each token these ten fields; in ISST-CoNLL only the fields ID, FORM, CPOSTAG, POSTAG, FEATS, HEAD and DEPREL are filled with information;
- data files are UTF-8 encoded (Unicode).

In what follows, the content of each field in ISST-CoNLL is described.

3.2 Field 1: ID

Token counter, starting at 1 for each new sentence.

3.3 Field 2: FORM

Word form or punctuation symbol.

3.4 Field 3: LEMMA

Lemma of the word form or punctuation symbol.

3.5 Field 4: CPOSTAG

This field contains for each token the coarse-grained part-of-speech tags which are based on the ILC/PAROLE tagset and are conformant to the EAGLES international standard. The table below documents the coarse-grained pos tags (14) used for ISST-CoNLL annotation.

Value:	**Description:**
A	adjective
B	adverb
C	conjunction
D	determiner
E	preposition
I	interjection
N	numeral
P	pronoun
PU	punctuation
R	article
S	noun
SA	abbreviation
V	verb
X	residual class

3.6 Field 5: POSTAG

This field contains for each token the fine-grained part-of-speech tags which are based on the ILC/PAROLE tagset and are conformant to the EAGLES international standard. The table below documents the fine-grained pos tags (28) used for ISST-CoNLL annotation.

Value	Description	Examples	Contexts of use
A	adjective	*bello, buono, pauroso, ottimo*	una **bella** passeggiata un **ottimo** attaccante una persona **paurosa**
AP	possessive adjective	*mio, tuo, nostro, loro*	a **mio** parere il **tuo** libro
B	adverb	*bene, fortemente, malissimo, domani*	arrivo **domani** sto **bene**
C	conjunction	*e, o, ma, mentre, quando*	i libri **e** i quaderni **quando** ho finito vengo
DD	demonstrative determiner	*questo, codesto, quello*	**questo** denaro **quella** famiglia
DE	exclamative determiner	*che, quale, quanto*	**che** disastro! **quale** catastrofe!
DI	indefinite determiner	*alcuno, certo, tale, parecchio, qualsiasi*	**alcune** telefonate **parecchi** giornali **qualsiasi** persona
DR	relative determiner	*cui, quale*	i **cui** libri
DT	interrogative determiner	*che, quale, quanto*	**che** cosa **quanta** strada **quale** formazione
E	preposition	*di, a, da, in, su, attraverso, verso, prima_di*	**a** casa **del** poeta **prima_di** giorno **verso** sera
I	interjection	*ahimè, beh, ecco, grazie*	**Beh**, che vuoi?
N	cardinal number	*uno, due, cento, mille, 28, 2000*	**due** partite **28** anni
NO	ordinal number	*primo, secondo, centesimo*	**secondo** posto
PD	demonstrative pronoun	*questo, quello, costui*	**quello** di Roma **costui** uccide
PI	indefinite pronoun	*chiunque, ognuno, molto*	**chiunque** venga i diritti di **ognuno**
PP	possessive pronoun	*mio, tuo, suo, loro, proprio*	il **mio** è qui più bella della **loro**
PQ	personal pronoun	*stressed: io, tu, egli, noi, voi* *unstressed: lo, la, mi, ci, vi*	**io** parto **lo** mangio
PR	relative pronoun	*che, cui, quale*	**ciò** che dice il **quale** afferma a **cui** parlo
PT	interrogative pronoun	*che, chi, quanto*	non so **chi** parta **quanto** costa? **che** ha fatto ieri?
PU	punctuation	*, ; . ? !*	mele**,** pere e banane**.**

Value	Description	Examples	Contexts of use
			cosa vuoi**?**
RD	determinative article	*il, lo, la, i, gli, le*	**il** libro **i** gatti
RI	indeterminative article	*uno, un, una*	**un** amico **una** bambina
S	common noun	*amico, insegnante, verità*	l'**amico** la **verità**
SA	abbreviation	*ndr, a.C., d.o.c., km*	30 **km** sesto secolo **a.C.**
SP	proper noun	*Monica, Pisa, Fiat, Sardegna*	**Monica** scrive
SW	foreign noun	*fazenda, mulieris dignitatem, weekend*	una **fazenda** in Brasile il prossimo **weekend**
V	verb	*mangio, avere, passato, camminando*	il peggio **è passato** **ho scritto** una lettera **vengo** domani
X	residual class	it includes formulae, unclassified words, alphabetic symbols and the like	distanziare di **43''** mi **piacce**

3.7 Field 6: FEATS

This field contains an unordered set of morph-syntactic features complementing the part of speech information. The tables which follow document:

1. the association between morpho-syntactic features and part of speech information (first table); features marked between square brackets are optionally specified;
2. the typology of features and their possible values (second table).

POS	Associated features
A	gen, num, [sup]
AP	gen, num
B	[sup]
E	[gen], [num]
DD, DE, DI, DR, DT	gen, num
N, NO	[gen], [num]
P, PD, PI, PP, PR, PT	gen, num
PQ	gen, num, per
RD, RI	gen, num
SA	gen, num
S, SP, SW	gen, num
V	[num], [per], mod, [tmp]

Tag	Feature values	Meaning	Example wordform pos feats
gen	M F N	masculine feminine underspecified tag subsuming M and F values	caso S **gen=M**|num=S giustizia S **gen=F**|num=S ospite S **gen=N**|num=S

num	S P N	singular plural underspecified tag subsuming S and P values	casa S gen=F|**num=S** donne S gen=F|**num=P** le PQ gen=F|**num=N**|per=3
mod	G F I C D M P	gerundive infinitive indicative subjunctive conditional imperative participle	dedicando V **mod=G** essere V **mod=F** trovava V num=S|per=3|**mod=I**|tmp=I possegga V num=S|per=3|**mod=C**|tmp=P sarebbero V num=P|per=3|**mod=D**|tmp=P cercate V num=P|per=2|**mod=M**|tmp=P rapita V gen=F|num=S|**mod=P**|tmp=R
per	1 2 3	first person second person third person	possiamo V num=P|**per=1**|mod=I|tmp=P sapete V num=P|**per=2**|mod=I|tmp=P le PQ gen=F|num=N|**per=3** vede V num=S|**per=3**|mod=I|tmp=P
tmp	P F I R	present tense future tense imperfect tense past tense	ha V num=S|per=3|mod=I|**tmp=P** sarà V num=S|per=3|mod=I|**tmp=F** trovava V num=S|per=3|mod=I|**tmp=I** rapita V gen=F|num=S|mod=P|**tmp=R**
sup	S	superlative	gravissimi A gen=M|num=P|**sup=S** benissimo B **sup=S**

3.8 Fields 7: HEAD

This field includes the non-projective head of current token, which is either a value of ID or zero ('0').

3.9 Field 8: DEPREL

This field includes the dependency relation to the non-projective-head, which is 'ROOT' when the value of HEAD is zero.

The table which follows documents the dependency tags (21) used for corpus annotation.

Tag	Relation type	Description	Examples
arg	argument	The most generic relation between a head and a subcategorized argument. Besides functional underspecification, this relation is always used to tag the syntactic relation between a verbal head and a non-subject clausal argument.	Il 63% dei francesi ha **imposto** al presidente **di** rinunciare alla sua bomba È giunto il **momento di** creare un’area denuclearizzata Le autorità hanno **annunciato che** il blitz è concluso Il **presidente dell’**assemblea

			nazionale
aux	auxiliary	The relation holding between a verb and its auxiliary.	Il corazziere **è stato** individuato Il corazziere è **stato individuato**
clit	clitic	The relation holding between a verbal head used in pronominal form and a clitic pronoun.	La sedia **si** è **rotta**
comp	complement	The most generic relation between a head and a complement, whether a modifier or a subcategorized argument. This underspecified dependency relation is particularly useful for those cases where it is difficult to draw a line between adjuncts and subcategorized elements. For instance, “comp” was resorted to for marking a) the relation between a head and a semantic argument syntactically realised as a modifier (as in the case of the agent as expressed in the passive construction), or b) the relation between the compared item and the comparative complement in comparative constructions.	Fu **assassinata da** un pazzo Il padre è stato **ucciso dai** banditi È più **interessante del** libro **Più di** quattrocento esemplari **Oggi come** allora
con	copulative conjunction	The relation “con” holds between a copulative conjunction in coordinate structures and the first conjunct which is taken to be the head of the whole coordinate structure.	Una ragazza **violentata e** sequestrata da due slavi **Gabriella e** Paolo sono partiti Hanno **riarmato ,** addestrato e preparato l’esercito Hanno **riarmato** , addestrato **e** preparato l’esercito **Scontri ,** assalti e centinaia di feriti **Scontri** , assalti **e** centinaia di feriti
concat	concatenation	The relation “concat” holds between tokens forming complex word forms (e.g. complex proper nouns, multi-word expressions and the like).	Il segretario di **De Michelis** L’enciclica “**Mulieris dignitatem**” La International **Public Sport** La **International Public** Sport

cong	in coordinate structures, the conjunct linked by copulative conjunction	The relation “cong” links the (second, third, …) conjuncts to the first conjunct which is taken as the head of the whole coordinate structure. “cong” is used in association with coordinating copulative conjunctions.	Una ragazza **violentata** e **sequestrata** da due slavi **Gabriella** e **Paolo** sono partiti Hanno **riarmato** , **addestrato** e preparato l’esercito Hanno **riarmato** , addestrato e **preparato** l’esercito **Scontri** , **assalti** e centinaia di feriti **Scontri** , assalti e **centinaia** di feriti
cong_sub	subordinative conjunction	The relation “cong_sub” holds between a subordinative conjunction and the verbal head of its clausal complement.	Ha detto **che** non **intendeva** fare nulla Le autorità hanno annunciato **che** il blitz è **concluso** Venne ucciso **mentre cercava** di difendere la ragazza
det	determiner	The relation holding between a nominal head and its determiner.	**Una sala** ha dovuto essere sgomberata Rilevata **la presenza** di gas
dis	disjunctive conjunction	The relation “dis” holds between a disjunctive conjunction in coordinate structures and the first conjunct which is taken to be the head of the whole coordinate structure.	Cassonetti dell’immondizia **rovesciati o** incendiati Partecipa **a** manifestazioni politiche **o** a dibattiti
disg	in coordinate structures, the conjunct linked by disjunctive conjunction	The relation “disg” links the (second, third, …) conjuncts to the first conjunct which is taken as the head of the whole coordinate structure. “disg” is used in association with coordinating disjunctive conjunctions.	Cassonetti dell’immondizia **rovesciati** o **incendiati** Partecipa **a** manifestazioni politiche o **a** dibattiti
mod	modifier	The relation “mod” holds between a head and its modifier, whether clausal or non-clausal.	I colori **sono sempre** gli stessi **Colori intensi** Si **valorizzano nella** luce mediterranea

			La **produzione di** Vietri **Trionfo** di Didoni **nei** 20 km di marcia Trionfo di Didoni nei **20 km** di marcia Trionfo di Didoni nei 20 **km di** marcia **Quando** urla non mi **piace** **Lavoravano** come volontari **per** costruire una centrale
mod_rel	relative modifier	The relation “mod_rel” holds between the verbal head of a relative clause and its nominal head in the higher clause. The same relation is also used in the case of free relatives to link the verbal head of the free relative and to the “chi” pronoun (which in its turn is directly linked to its governor)	**Box** che è stato **trovato** nel pomeriggio Gli utilizzatori del **box** dove sarebbe **avvenuta** la violenza Non è mai stato accertato **chi volle** la sua morte
modal	modal verb	The relation holding between a verbal head and a modal verb.	Una sala ha **dovuto** essere **sgomberata**
obl	oblique complement	The relation “obl” holds between a verbal head and a subcategorized non-direct, non indirect and non-clausal complement.	Si è **trasformato in** gas Era **uscito di** casa alle 10 Sono **volati nel** burrone
ogg_d	direct object	The relation “ogg_d” holds between a verbal head and its direct object (always non-clausal).	**Hanno** un **modo** di ragionare rozzo **Centellinando** le **informazioni** È giunto il momento di **creare** un’**area** denuclearizzata **Rilevata** la **presenza** di gas
ogg_i	indirect object	The relation “ogg_i” holds between a verbal or nominal head and the indirect object, i.e. the complement expressing the recipient or beneficiary of the action expressed by the verb or the noun.	**Al** magistrato **ripetevano** **Mi** ha **elargito** uno sguardo Un **contributo alla** lotta contro la criminalità

pred	predicative complement	The relation “pred” holds between a head and a predicative complement, be it subject or object predicative.	L’incontro è **stato fatale** Questo **è** il **messaggio** finale Lo hanno **chiamato Giovanni** Li **considera** troppo **abbaglianti**
prep	preposition	The relation “prep” holds between a prepositional head and its complement, whether clausal or non-clausal.	Un contributo **alla lotta** contro la criminalità Un contributo alla lotta **contro** la **criminalità** Hanno un modo **di ragionare** rozzo **Prima_di partire** ho telefonato
punc	punctuation	The relation “punc” holds between a word token and a punctuation mark.	Teatro della **tragedia ,** ... Una **polemica :** **“ Blitz** concluso” “ Blitz **concluso ”**
sogg	subject	The relation “sogg” holds between a verb and its subject: 1. “sogg” refers to the superficial subject of a verb, regardless of the latter being used in the active or passive voice 2. “sogg” is also used to mark clausal subjects 3. with pro-drop languages such as Italian, when the subject is not overtly realised the annotation does not include an explicit subject relation: the morphosyntactic features, indicating person, number and possibly gender of the subject, can be recovered from the inflectional features associated with the verb.	1. il **testimone** ha **parlato** subito i **missionari** erano stati **rapiti** la mattina presto le **vittime seguivano** gli aiuti 2. **è** opportuno **dire** due parole **è** difficile **che** la gente se ne vada presto **sarà** difficile **negare** 3. **arrivo** domani **colpiscono** le statue

3.10 Field 9: PHEAD

Projective head of current token, which in ISST-CoNLL is always an underscore.

3.11 Field 10: PDEPREL

Dependency relation to projective head, which is always an underscore.

Acknowledgements

Isidoro Barraco and Patrizia Topi did most of the work of writing conversion scripts and revising tags. Kiril Ribarov, Alessandro Lenci and Giuseppe Attardi contributed useful discussions on critical issues.

References

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte (2003a) “Building the Italian Syntactic-Semantic Treebank”, in Anne Abeillé (ed.), *Building and using Parsed Corpora*, Language and Speech series, Kluwer, Dordrecht, pp. 189-210.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Vito Pirrelli, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte (2003b) “The syntactic-semantic treebank of Italian. An overview”, *Linguistica Computazionale* XVI-XVII, pp. 461-492