# EVALITA 2011
# Super Sense Tagging Task

*Stefano Dei Rossi, Giulia Di Pietro, Maria Simi*
Dipartimento di Informatica, Università di Pisa

## 1 Task description

Super-sense tagging (SST) is a Natural Language Processing task that consists of annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, within a general semantic taxonomy defined by the WordNet lexicographer classes (called super-senses) [1]. SST can be considered as a task half-way between Named-Entity Recognition (NER) and Word Sense Disambiguation (WSD): it is an extension of NER, since it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD, that deals with very specific senses.

The task is to predict an appropriate super sense for each token or multiword expression. For instance expressions such as *"Croce Rossa"*, "Fiona May" and "10 dicembre 1975" are considered as single entities
Modal and support verbs are not annotated since they do not entail any semantics.

The 45 Super-Sense categories (3 used for adjectives, 25 for distinguish nouns, 15 for verbs and one for adverbs) are shown in the following table. More details are reported in Appendix A.

| Id | SuperSense | Description | Examples |
|---|---|---|---|
| 00 | adj.all | all adjective clusters, used for all simple adjectives | grande, bello, simpatico |
| 01 | adj.pert | relational adjectives (pertainyms), adjectives that are related with nouns | scolastico, marittimo |
| 02 | adv.all | all adverb | anche, sempre, dove |
| 03 | noun.Tops | unique beginner for nouns, nouns that appear as super senses | animale, gruppo, tempo |
| 04 | noun.act | nouns denoting acts or actions | corsa, |
| 05 | noun.animal | nouns denoting animals | cane, gorilla, coniglio |
| 06 | noun.artifact | nouns denoting man-made objects | edificio, fontana, bomba |
| 07 | noun.attribute | nouns denoting attributes of people and objects | ricchezza, pigrizia, eleganza |
| 08 | noun.body | nouns denoting body parts | braccio, occhio, naso |
| 09 | noun.cognition | nouns denoting cognitive processes and contents | Pensiero, sogno, conoscenza |
| 10 | noun.communication | nouns denoting communicative processes and contents | libro, licenza, discussione |
| 11 | noun.event | nouns denoting natural events | trionfo, incidente |
| 12 | noun.feeling | nouns denoting feelings and emotions | delusione, paura, desiderio |
| 13 | noun.food | nouns denoting foods and drinks | miele, aranciata, pizza, cena, merenda |

| | | | |
|---|---|---|---|
| 14 | noun.group | nouns denoting groupings of people or objects | Chiesa, ONU, Mediaset. Italia,Francia, Germania |
| 15 | noun.location | nouns denoting spatial position | Pisa, Roma,via, piazza |
| 16 | noun.motive | nouns denoting goals | ragione, causa, motivo |
| 17 | noun.object | nouns denoting natural objects (not man-made) | pietra, mare, montagna |
| 18 | noun.person | nouns denoting people | Giovanni, Rossi |
| 19 | noun.phenomenon | nouns denoting natural phenomena | nebbia, fulmine, perturbazione |
| 20 | noun.plant | nouns denoting plants | pino, polline, basilico |
| 21 | noun.possession | nouns denoting possession and transfer of possession | finanziamento, tassa, miliardi. |
| 22 | noun.process | nouns denoting natural processes | declino, sviluppo tramonto |
| 23 | noun.quantity | nouns denoting quantities and units of measure | metri, dollari |
| 24 | noun.relation | nouns denoting relations between people or things or ideas | per cento, parte, est, ovest |
| 25 | noun.shape | nouns denoting two and three dimensional shapes | colonna, piano, curva |
| 26 | noun.state | nouns denoting stable states of affairs | morte, crisi, pace |
| 27 | noun.substance | nouns denoting substances | oro, gas, pasta |
| 28 | noun.time | nouns denoting time and temporal relations | notte, settembre, ore |
| 29 | verb.body | verbs of grooming, dressing and bodily care | dormire, respirare |
| 30 | verb.change | verbs of size, temperature change, intensifying, etc. | accendere, chiudere |
| 31 | verb.cognition | verbs of thinking, judging, analyzing, doubting | immaginare, dubitare, sperare |
| 32 | verb.communication | verbs of telling, asking, ordering, singing | parlare, cantare, leggere |
| 33 | verb.competition | verbs of fighting, athletic activities | vincere, gareggiare or espugnare, sparare |
| 34 | verb.consumption | verbs of eating and drinking | mangiare, sorseggiare, fumare |
| 35 | verb.contact | verbs of touching, hitting, tying, digging | avvolgere, sfiorare |
| 36 | verb.creation | verbs of sewing, baking, painting, performing | costruire, distruggere, dipingere, suonare |
| 37 | verb.emotion | verbs of feeling | esaltare, temere |
| 38 | verb.motion | verbs of walking, flying, swimming | camminare, volare, muovere |
| 39 | verb.perception | verbs of seeing, hearing, feeling | vedere, sentire |
| 40 | verb.possession | verbs of buying, selling, owning | finanziare, pagare, investire |
| 41 | verb.social | verbs of political and social | presentare, organizzare, |

| | | | |
|---|---|---|---|
| | | activities and events | emarginare |
| 42 | verb.stative | verbs of being, having, spatial relations | rimanere, mantenere, esistere |
| 43 | verb.weather | verbs of raining, snowing, thawing, thundering | piovere, nevicare, tuonare |
| 44 | adj.ppl | participial adjectives | Preoccupante ?? |

### 1.1 Substaks

**Closed subtask**
In the closed subtask, we want to measure the accuracy in SS tagging, when only the corpus provided for training is used.

**Open subtask**
In the open subtask participants will be free to use any external resource in addition to the corpus provided for training; for example, instances of WordNet as well as other lexical or semantic resources.

## 2 Corpus description

### 2.1 Source of training data

A corpus for Super-sense tagging was created starting from the Italian Syntactic-Semantic Treebank (ISST) [2] by a semi-automatic correction and conversion process, followed by manual revision. This process is detailed in [3].
ISST-SST (about 300,000 tokens) will be made available for the task and for research purposes. A portion of about 276,000 tokens will be used for training and development.
The evaluation will be performed on a smaller corpus obtained from a held-out portion of ISST-SST (about 30,000 tokens) and a portion of the Italian Wikipedia (about 20,000 additional tokens).
The creation of ISST-SST was initiated as part of the project SemaWiki (Text Analytics and Natural Language processing - TANL) [4], a collaboration between the University of Pisa and the Institute for Computational Linguistics of CNR.

### 2.2 Training corpus statistics

The training corpus consists in about 276,000 word forms divided into 11,342 sentences.

| | |
|---|---|
| #documents | 430 |
| #sentences | 11,342 |
| #tokens | 276,423 |
| #Annotated tokens | 135,738 |

### 2.3 Data format

Data adheres to the following rules:
1. Characters are UTF-8 encoded (Unicode).
2. Data files are organized in documents. Each document starts with the line
   `-DOCSTART- -X- O O`
3. A document contains sentences separated by an empty line.
4. A sentence consists of a sequence of tokens, one token per line.

5. A token consists of four fields described in the table below. Fields are separated by one tab character.
6. SST tags can span several tokens and use the IOB notation: labels are prefixed with "B" for begin, "I" for inside, and "O", outside any label. This notation is typical in several CoNLL tagging tasks.

| Field Name | Description |
|---|---|
| Form | Word form or punctuation symbol |
| Lemma | Word lemma or punctuation symbol |
| PoS | Part-of-speech tag, with morphological features, based on the TANL tagset. |
| Super Sense Tag | Super Sense tag in IOB notation |

**Example**

```
VENARIA venaria SP      B-noun.location
(       ,       FF      O
Torino  torino  SP      B-noun.location
)       )       FB      O
-       -       FC      O
Un      un      RIms    O
incendio        incendio        Sms     B-noun.event
,       ,       FF      O
che     che     PRnn    O
si      si      PC3nn   O
sarebbe essere  VAd3s   O
sviluppato      sviluppare      Vpsms   B-verb.creation
per     per     E       O
cause   causa   Sfp     B-noun.motive
accidentali     accidentale     Anp     B-adj.all
,       ,       FF      O
ha      avere   VAip3s  O
gravemente      gravemente      B       B-adv.all
danneggiato     danneggiare     Vpsms   B-verb.change
a       a       E       O
Fiano   fiano   SP      B-noun.location
(       ,       FF      O
Torino  torino  SP      B-noun.location
)       )       FB      O
,       ,       FF      O
uno     uno     RIms    O
chalet  chalet  Smn     B-noun.artifact
di      di      E       O
proprietà       proprietà       Sfn     B-noun.possession
di      di      E       O
Umberto umberto SP      B-noun.person
Agnelli agnelli SP      I-noun.person
```

### 2.4 The TANL tagset

The TANL morpho-syntactic annotation schema was developed in the SemaWiki project:
http://medialab.di.unipi.it/wiki/SemaWiki
For a description of the Tanl tagset for morph-syntactic annotation see:
http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset
For annotation guidelines: http://medialab.di.unipi.it/wiki/Annotation_guidelines

### 2.5 Copyright and license

ISST-SST is copyrighted material which can be used for research purposes only and which cannot be distributed in any original or modified form. Participants will be requested to agree on these terms and conditions upon downloading the resource.

## 3 Resource download

The following resources:

- ISST-SST Training Corpus (1st version)
- SST Tagging Accuracy Evaluator

can be download from the site: http://medialab.di.unipi.it/evalita2011/

## 4 Submission details

Participants should submit their results by October 14th, midnight Italian time.

Runs must be sent to the organizers address, evalita@di.unipi.it, as a file in the same format as the Training Corpus, named as:

<team>_SST_<Open|Closed>_<run>

<team>: a short name for the team, without special characters

<Open|Closed>: Open or Closed subtask

The assessment of the submitted runs will be sent to the participants by October 28th, 2011, together with the gold-standard version of the test data.

## 5 Evaluation

The evaluation metrics will be:

- *tagging accuracy*, i.e. the percentage of correctly classified tokens with respect to the total number of tokens.
- *precision*, the percentage of correct positive predictions over the total number of positive predictions by the system
- *recall*, the percentage of correct positive predictions by the system over the expected predictions
- *F1-measure*, the weighted harmonic mean of precision and recall

An evaluation script, adapted from the CoNLL2000 shared task on chunking, is made available for evaluation purposes. The Perl evaluation script conlleval.pl can be used as follows:

```
conlleval.pl -g <gold-file> -s <sys-output>
```

Participants are required to provide a brief description of their system and a full notebook paper describing their experiments, in particular the techniques and the resources used, and presenting an analysis of the results.

## 6 Contacts

Stefano Dei Rossi <deirossi@di.unipi.it>
Maria Simi <simi@di.unipi.it>

*Dipartimento di Informatica, Università di Pisa*
*Largo B. Pontecorvo, 3*
*I-56127 Pisa*
*Italy*
*Phone: (+39) 050 2212758*
*Fax: (+39) 050 22127266*

## 7 Acknowledgements

Giuseppe Attardi, Alessandro Lenci, Simonetta Montemagni.

## 8 References

[1] C. Fellbaum (Ed.) (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge.

[2] S. Montemagni, et al. 2003. Building the Italian Syntactic-Semantic Treebank. In Abeillé (ed.), Building and using Parsed Corpora, Language and Speech series, Kluwer, Dordrecht, 189–210.

[3] G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi, A Resource and Tool for Super-sense Tagging of Italian Texts, Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), Malta, 17-23 May 2010.

[4] G. Attardi et al. 2008. Tanl (Text Analytics and Natural Language processing). Project Analisi di Testi per il Semantic Web e il Question Answering, http://medialab.di.unipi.it/wiki/SemaWiki.

APPENDIX A

**0 - adj.all**

This tag is used for all simple adjectives, such as *"grande", "bello", "simpatico"*.

**1 - adj.pert**

This tag is used for all adjectives that are related with nouns, such as *"scolastico", "marittimo", "soleggiato"*.

**2 - adv.all**

This tag is used for all adverbs, such as *"anche", "sempre", "dove"*.

**3 - noun.Tops**

This tag is used for those nouns that appear as super sense, such as *"animale", "gruppo", "tempo"*.

**4 - noun.act**

This tag is used for all those nouns that denote an action, such as *"corsa", "incontro", "sciopero"*.

**5 - noun.animal**

This tag is used for all nouns of animals, such as *"cane"*, *"gorilla"*, *"coniglio"*.

**6 - noun.artifact**

This tag is used for all man-made objects, such as *"edificio"*, *"fontana", "bomba"*.

**7 - noun.attribute**

This tag is used for all nouns that denote attributes of people, such as *"serietà", "eleganza", "pigrizia"*.

**8 - noun.body**

This tag is used for all body parts, such as*"braccio"*, *"occhio", "cuore"*.

**9 - noun.cognition**

This tag is used to identify all nouns related to cognitive (or mental) processes, such as *"pensiero", "sogno", "conoscenza"*.

**10 - noun.communication**

This tag is used for all nouns that denote both objects that allow communication, such as *"libro"*, *"film"*, *"licenza"*, and nouns that denote comunicative processes, such as *"discussione", "chiarimento", "proposta"*.

**11 - noun.event**

This tag is used to denote all nouns of event, such as *"trionfo"* or *"incidente"*.

**12 - noun.feeling**

This tag is used to identify all emotions and feelings, such as *"delusione", "paura", "desiderio"*.

**13 - noun.food**

This tag is used to denote both all nouns of food, such as *"miele"*, *"aranciata", "pizza"*, and the meals in which they are consumed, such as *"cena"* o *"merenda"*.

**14 - noun.group**

This tag is used to denote all the nouns that refer to associations or organization, groups or communities, such as *"church"*, *"ONU", "Mediaset"*. This tag is also used to denote, football team, such as *"Italia", "Francia", "Germania"*.

**15 - noun.location**

This tag is used to denote nouns of cities or places, such as *"Pisa"*, *"Roma"* or *"via"*, *"piazza"*.

**16 - noun.motive**

This tag is used to denote all those nouns that refer to a purpose, such as *"ragione"*, *"causa"*, *"motivo"*.

**17 - noun.object**

This tag is used to denote all natural objects, such as *"pietra"*, *"mare", "montagna"*, but just if they are used as objects. For instance, in this sentence

"Abbiamo scalato la montagna più alta del mondo"

*montagna* is annoted as *noun.object*, but on the other hand, in this sentence
"Sono andato in montagna"
montagna is annoted as noun.location.

**18 - noun.person**

This tag is used to denote first and last name of persons, such as *"Giovanni"* or *"Rossi"*. This tag is also used to denote

**19 - noun.phenomenon**

This tag is used for nouns that denote natural phenomenons, such as *"nebbia"*, *"fulmine", "perturbazione"*.

**20 - noun.plant**

This tag is used for all nouns of plants, such as *"pino"*, *"polline"*, *"basilico"*. Is also used for the nouns of vegetables, but not used in context of eating.

**21 - noun.possession**

This tag is used for all nouns of pessession, such as *"finanziamento"*, *"tassa"*, and also for nouns of quantity of cash, such as *"miliardi"*.

**22 - noun.process**

This tag is used to denote all nouns that express the gowing up of a process, such as *"declino", "sviluppo",* and also natural processes, such as *"tramonto"*.

**23 - noun.quantity**

This tag is used for all nouns that denote a quantity or units, such as numbers or *"metri"* and *"dollari"*.

**24 - noun.relation**

This tag is used to all nouns that refer to a part of something, such as *"per cento"* or *"parte",* but also *"est"* or *"ovest"*, because they refer to a single part of something (for instance, a *eastern* part fo the world).

**25 - noun.shape**

This tag is used to all all nouns that denote a objects that have a particolar shape, such as *"colonna"*, *"piano"*, *"curva"*.

**26 - noun.state**

This tag is used to all nouns that denote a state of persons or situations, such as *"morte", "crisi", "pace"*.

**27 - noun.substance**

This tag is used to all nouns that denote a substance, such as *"oro"*, *"gas", "pasta"*.

**28 - noun.time**

This tag is used to all nouns that express time, such as *"notte", "settembre"*, *"ore"*.

**29 - verb.body**

This tag is used to all verbs that express actions of body, such as *"dormire"*, *"respirare"*.

**30 - verb.change**

This tag is used to all verbs that express a change of something, such as "a*ccendere"*, *"chiudere"*.

**31 - verb.cognition**

This tag is used to all verbs that express actions that involve the mind, such as *"immaginare", "dubitare", "sperare"*.

**32 - verb.communication**

This tag is used to all communication verbs, such as *"parlare"*, *"cantare"*, *"leggere"*.

**33 - verb.competition**

This tag is used to all those verbs of both sports competition and hostility, such as *"vincere", "gareggiare"* or *"espugnare", "sparare"*.

**34 - verb.consumption**

This tag is used to all verb that express action of eating, drinking or, more generally, consumption of something, such as *"mangiare"*, *"sorseggiare"* or *"fumare"*.

**35 - verb.contact**

This tag is used for all verbs that denote contact, such as *"avvolgere"*, *"sfiorare"*.

**36 - verb.creation**

This tag is used to all verbs that express action of creation – or destruction –,such as *"costruire"* and *"distruggere"*, but also verbs about creative processes, such as *"dipingere" or "suonare"*.

**37 - verb.emotion**

This tag is used to all verbs of emotion or feelings, such as *"esaltare"*, *"temere"*.

**38 - verb.motion**

This tag is used for all verbs that express different type of moving, such as *"camminare"*, *"volare"* o *"muovere"*.

**39 - verb.perception**

This tag is used to all verbs about perception, such as *"vedere"*, *"sentire"*.

**40 - verb.possession**

This tag is used for all verbs that express Exchange of possessions, such as *"finanziare"*, *"pagare"*, *"investire"*.

**41 - verb.social**

This tag is used for all verbs used to express social action, such as *"presentare"*, *"organizzare"*, *"emarginare"*.

**42 - verb.stative**

This tag is used to all verbs that express a state that not change, such as *"rimanere", "mantenere", "esistere"*.

**43 - verb.weather**

This tag is used to all verbs that express weather situations, such as *"piovere"*, *"nevicare"*, *"tuonare"*.

**44 - adj.ppl**

This tag is used to denote all adjectives participials, i.e. those adjectives that have the same form of a participle, but they are not related with some verb, such as *"preoccupante"*.