XML Format of the Train and Test Sets for the Italian Lexical Sample Task at Senseval-3 (new version)

The trial **trial set** for the Italian Lexical Sample task included the words *intervento* (noun) and *scartare* (verb), together with the DTD file *s-3lex-sample.dtd*.
In order to map the ambiguous word in the PoS-tagged snippets, we have introduced a little change in the XML format of both **train and test sets**, and we consequently modified the DTD.
So, please use the new DTD file *s-3lex-sample-v2.dtd*, that is included in all the 5 tar.zipped folders provided to participants. We provide both labelled and unlabelled examples, as follows:

LEMMA	PoS	TRAIN SET		TEST SET
		# LABELLED INSTANCES (both PoStagged and not-PoStagged versions are available)	# UNLABELLED INSTANCES (only not-PoStagged version is available)	# INSTANCES (both PoStagged and not-PoStagged versions are available)
acuto	adjective	102	780	48
attacco	noun	112	1751	53
canale	noun	120	1881	58
carattere	noun	111	1731	52
chiaro	adjective	113	1751	52
classe	noun	148	2271	69
comodo	adjective	102	928	48
complesso	noun	86	1371	41
corona	noun	132	988	61
corso	noun	124	1931	59
decidere	verb	111	1751	54
discendere	verb	92	544	43
elemento	noun	117	1821	55
esecuzione	noun	96	1520	46
esposizione	noun	102	1600	48
espressione	noun	104	1650	51
giocare	verb	142	2201	68
giro	noun	133	2051	62
gonfio	adjective	119	620	56
gruppo	noun	109	1701	51
guida	noun	117	1821	55
guidare	verb	112	1751	53
lettera	noun	81	1301	39
massa	noun	122	1901	58
modesto	adjective	102	1600	48
naturale	adjective	153	2350	72
nota	noun	112	1751	53
operazione	noun	111	1741	53
ordine	noun	122	1901	58
partito	noun	71	1151	34
perdere	verb	184	2801	86
programma	noun	128	1971	59
rendere	verb	122	1911	59
resistenza	noun	111	1750	54
rete	noun	174	2651	81
scalare	verb	91	857	44
scorrere	verb	112	1497	53
scuro	adjective	81	1300	39
segnare	verb	121	1901	59
semplice	adjective	122	1901	58
sicuro	adjective	126	1951	59
teso	adjective	91	1384	44
vertice	noun	101	1601	49
vincere	verb	112	1751	53
vista	noun	91	1451	44

In the Italian Lexical Sample Task we will consider 45 words (25 nouns, 10 adjectives and 10 verbs).
The number of instances for each word varies according to the number of senses that we have taken into consideration:

- # instances for each word in the test set = {1/3 * [75 + (15*#senses)+(7*#multiwords)]}
- # labelled instances for each word in the train set = {2/3 * [75 + (15*#senses)+(7*#multiwords)]}
- # unlabelled instances for each word in the train set = {about 10 * [75 + (15*#senses)+(7*#multiwords)] + 100}

For six words we could not collect automatically all the unlabelled train examples we expected (i.e. 10 times the examples we extracted manually). So, the words “acuto”, “comodo”, “corona”, “discendere”, “gonfio”, “scalare” have less train examples than the other ones.

Participants are provided with 5 tar.zipped folders:
a.1) **test-postagged.zip** (1.9M)
a.2) **test-not-postagged.zip** (0.4M)
b.1) **train-postagged.zip** (4M)
b.2) **train-not-postagged.zip** (1M)
b.3) **train-automatic-not-postagged.zip** (15M)

Each folder contains the examples for all the 45 words we considered:
a.1) is the test set containing PoS-tagging information;
a.2) is the test set without PoS-tagging information;
b.1) is the train set containing PoS-tagging information: all instances have been manually tagged;
b.2) is the train set without PoS-tagging information: all instances have been manually tagged;
b.3) is another part of the train set: all instances (about 10 times a+b) have been automatically extracted from the MEANING corpus and were neither tagged nor PoS-tagged.

Participants can process the PoS-tagged or the not-PoS-tagged versions of the data, depending on their needs.

Examples:

Here are three samples from *intervento.xml* ("intervento" appears neither in the test nor in the train set):

1. The first example (id="intervento.1") shows the PoS tagging information (within the tag <postagging>). As can be noticed, the word that must be disambiguated is identified in both <context> and <postagging>.
   For brevity's sake, in the two following samples the PoS of each token has been omitted.
   We used the TnT Part-of-Speech Tagger developed by Thorsten Brants. The PoS tagger is described in: Thorsten Brants, "TnT – A Statistical Part-of-Speech Tagger", in Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, April 29 - May 3, 2000, Seattle, WA.
2. In the second instance (id="intervento.5") we have a multiword: the satellite is directly connected to the head. The tag <head> has an attribute sats, whose value is the same of the attribute id in the tag <sat>.
3. The third example (id="intervento.70") is not labelled. As you can notice, there is no <answer> tag. The train set that participants will receive will be made up of both labelled and unlabelled examples (around ten times more).

Sample 1:

```
<corpus type="trial" lang="italian">
 <lemma item="intervento" pos="n">

        <instance id="intervento.1">
                <context>
Potrebbe essere partita dalle agocannule usate per l'<head>intervento</head> di
liposuzione al ginocchio la terribile infezione che ha ridotto in fin di vita due donne
e in gravi condizioni una terza.
                </context>
                <postagging>
                 <word id="0" pos="VI">
                  <token>Potrebbe</token>
                  <lemmas>potere</lemmas>
                 </word>
                 <word id="1" pos="VFY">
                  <token>essere</token>
                 </word>
                 <word id="2" pos="VSP">
                  <token>partita</token>
                  <lemmas>partire</lemmas>
                 </word>
                 <word id="3" pos="EP">
                  <token>dalle</token>
                  <lemmas>da+le</lemmas>
                 </word>
                 <word id="4" pos="SP">
                  <token>agocannule</token>
                 </word>
                 <word id="5" pos="AP">
                  <token>usate</token>
                  <lemmas>usato</lemmas>
                 </word>
                 <word id="6" pos="E">
                  <token>per</token>
                  <lemmas>per</lemmas>
                 </word>
                 <word id="7" pos="RS">
                  <token>l'</token>
                  <lemmas>det</lemmas>
                 </word>
                 <word id="8" pos="SS" annotated="head">
                  <token>intervento</token>
                  <lemmas>intervento</lemmas>
                 </word>
                 <word id="9" pos="E">
                  <token>di</token>
                  <lemmas>di</lemmas>
                 </word>
                 <word id="10" pos="SS">
                  <token>liposuzione</token>
                 </word>
                 <word id="11" pos="ES">
                  <token>al</token>
                  <lemmas>a+il</lemmas>
                 </word>
                 <word id="12" pos="SS">
                  <token>ginocchio</token>
                  <lemmas>ginocchio</lemmas>
```

```
        </word>
        <word id="13" pos="RS">
         <token>la</token>
         <lemmas>det</lemmas>
        </word>
        <word id="14" pos="AS">
         <token>terribile</token>
         <lemmas>terribile</lemmas>
        </word>
        <word id="15" pos="SS">
         <token>infezione</token>
         <lemmas>infezione</lemmas>
        </word>
        <word id="16" pos="CCHE">
         <token>che</token>
         <lemmas>che</lemmas>
        </word>
        <word id="17" pos="VIY">
         <token>ha</token>
         <lemmas>avere</lemmas>
        </word>
        <word id="18" pos="VSP">
         <token>ridotto</token>
         <lemmas>ridurre</lemmas>
        </word>
        <word id="19" pos="E">
         <token>in</token>
         <lemmas>in</lemmas>
        </word>
        <word id="20" pos="SS">
         <token>fin</token>
        </word>
        <word id="21" pos="E">
         <token>di</token>
         <lemmas>di</lemmas>
        </word>
        <word id="22" pos="SS">
         <token>vita</token>
         <lemmas>vita</lemmas>
        </word>
        <word id="23" pos="N">
         <token>due</token>
        </word>
        <word id="24" pos="SP">
         <token>donne</token>
         <lemmas>donna</lemmas>
        </word>
        <word id="25" pos="C">
         <token>e</token>
         <lemmas>e</lemmas>
        </word>
        <word id="26" pos="E">
         <token>in</token>
         <lemmas>in</lemmas>
        </word>
        <word id="27" pos="AP">
         <token>gravi</token>
         <lemmas>grave</lemmas>
        </word>
        <word id="28" pos="SP">
         <token>condizioni</token>
         <lemmas>condizione</lemmas>
        </word>
        <word id="29" pos="RS">
         <token>una</token>
         <lemmas>indet</lemmas>
        </word>
        <word id="30" pos="SS">
         <token>terza</token>
         <lemmas>terza</lemmas>
        </word>
        <word id="31" pos="XPS">
         <token>.</token>
        </word>
       </postagging>
<answer id="intervento.1" offset="n#00436664" sense="1"></answer>
</instance>
```

Sample 2:

<instance id="intervento.5">
<context>
Nel '92 avevo un corpo trasformato e per questo mi rivolsi al professor Conconi. Lui mi indirizzò da un endocrinologo. Da allora sono costantemente sotto terapia: non sono dopata". Soprani ha anche insistito su quell'improvviso <head sats="intervento_chirurgico.5"> intervento </head> <sat id="intervento_chirurgico.5"> chirurgico </sat> per dolori addominali, avvenuto a Ferrara nel novembre '94. Ma l' attività del pm non si esaurisce col file di Conconi: in Danimarca ha acquisito dati su Bjarne Riis, vincitore del Tour '96, a Boston uno studio che abbassa il valore dell'ematocrito ritenuto normale.
</context>
<postagging>

[...]

</postagging>
<answer id="intervento.5" offset="n#00436664" sense="1"></answer>
</instance>

Sample 3:

<instance id="intervento.70">
<context>
Ad esempio il rifiuto di Arafat come interlocutore, da parte del governo israeliano. Un <head>intervento</head> ancor più severo dell'esercito. E la possibilità, in prospettiva, di un gabinetto di emergenza: vale a dire una grande coalizione destra-sinistra, che non consentirebbe più di sperare, nel futuro scrutabile, in una ripresa, in un riaggiustamento del frantumato processo di pace.
</context>
<postagging>

[...]

</postagging>
</instance>

XML Format:

The following table describes the tags and attributes of the XML format for the Italian Lexical Sample task data.
The attribute '`annotated`' in the tag `<word>` did not appear in the trial set format, and has been added to the train and test sets.

TAG	DESCRIPTION	ATTRIBUTE	VALUE
corpus	Participants in the Italian Lexical Sample task are provided with three corpora: trial, train and test.	**type**	"trial", "train" or "test"
		lang	"italian"
lemma	It contains the instances of the word. The Italian Lexical Sample train and test sets will have 45 lemmata.	**item**	Lemmatized form of the words to be disambiguated.
		pos	PoS of the word to be disambiguated: "n" for nouns, "a" for adjectives and "v" for verbs.
instance	Numbered examples.	**id**	From "1" to n, where n depends on the number of senses we considered.
context	Text snippet that contains the word to disambiguate.	**/**	/
head	It contains the token to disambuguate.	**sats**	This attribute appears only with multiwords, and its value is the entire multiword followed by the number of the instance (like the attribute `id` of the tag `<sat>`).
sat	It contains the satellite of the multiwords.	**id**	Same as the attribute `sats` of the tag `<head>`.
postagging	It returns the output of the TnT postagger, for each text snippet.	**/**	/
word	It keeps track of each word that appears in the text snippet.	**id**	From "1" to n, where n is the number of words in the text snippet.
		pos	A tag from the TnT tagset (see Appendix)
		annotated	This attribute identifies the word that must be disambiguated. Its value can be "head" or "sat".
token	Each single token from the text snippet.	**/**	/
lemmas	Lemmatized form of each token. This information is not always available. When there are more lemmata per a token, they are separated by a comma	**/**	/
answer	It returns the correct meaning of the word to disambiguate. The content of this tag is always empty, and the correct answer is expressed by the attributes.	**id**	Number of the instance it is referred to.
		offset	It return the correct synset offset drawn from "MultiWordNet-for-Senseval3", which is the sense inventory for the Italian Lexical Sample task.
		sense	It returns the correct sense number, drawn from "MultiWordNet-for-Senseval3".

Appendix: TnT PoS-Tagger Tagset:

Tag	Description	Example
XPS	Punctuation	. ; : ? !
XPW	Comma	,
XPB	Brackets	( )
XPO	quotation mark, ellipsis, hyphen	`` '' - ... _
N	Number	1, 1999, '76, sei, sesto
RS	singular article	il, l', la, un, una
RP	plural article	i, gli, gl', le
AS	singular qual. adj.	vera, grandissimo, migliore
AP	plural qual. Adj.	vere, grandissimi, maggiori
AN	qual. adj. neutral for number	rosa, più, super, antincendio
D	singular det. Adj.	quello, alcuna, mio, quale?
DP	plural det. Adj.	quelli, alcune, miei, quali?
DN	det. adj. neutral for number	qualsiasi
E	simple preposition	di, a, dopo, fino, nonostante
ES	singular articulated preposition	dal, sulla, nello
EP	plural articulated preposition	dalle, sulle, negli, nei, ai
B	Adverb	molto, invece, esattamente
C	Conjunction	e, ma, bensı, sia, perché
CCHE	Che	che
CCHI	Chi	chi
CADV	connettivo avverbiale	come, dove, quando
PS	singular pronoun	ciascuna, lo, mio
PP	plural pronoun	costoro, esse, nostri, loro
PN	pronoun neutral for number	ci, cui, sé
SS	singular noun	aereo, formula
SP	plural noun	aerei, formule
SN	noun neutral for number	attività, business, novità
SPN	proper noun	Alfredo, Ford, Piombino
QNS	singular relative pronoun	quanto, quanta
QNP	plural relative pronoun	quanti, quante
YA	Acronym	ANSA, CEE, ONU
YF	foreign term	city, fiesta, Papier
I	Interjection	oh!
VI	main verb, ind., subjunctive, cond.	vedo, giungano, saprei
VIY	aux. verb, ind., subjunctive, cond.	ho, sia, avrebbe
VF	main verb, inf.	arrivare, vedere
VFY	aux. verb, inf.	avere, essere
VSP	main verb, past part., singular	acquisito, interrotto
VSPY	aux. verb, past part., singular	avente, stato, stata
VPP	main verb, past part., plural	arrivati
VPPY	aux. verb, past part., plural	state
VG	main verb, gerund	cantando, ringraziando
VGY	aux. verb, gerund	avendo, essendo
VM	main verb, imperative	cercate, leggi
VMY	aux. verb, imperative	sia, abbia
+E	Clitic	ne, ci