STiL – Studies in Linguistics Vol. 2, 2008
CISCL Working Papers

A Bilingual Treebank (ITA-LIS) suitable for Machine Translation: what Cartography and Minimalism teach us

Cristiano Chesi*, Gianluca Lebani+,
Margherita Pallottino*
*CISCL - University of Siena, +CIMeC - University of Trento
chesi@media.unisi.it, gianluca.lebani@unitn.it, margherita.pallottino@yahoo.com

In this paper we describe some technical and theoretical aspects related to a manually aligned bilingual treebank Italian (ITA) – Italian Sign Language (LIS) provided with both constituency and dependency annotation (Siena University Treebank, SUT). We briefly discuss the linguistic rationale behind the feature set and the dependency/constituency structure we adopted. Moreover we discuss the tool we used to annotate, semi-automatically, the treebank that, in the end, will be evaluated qualitatively with respect to a specific Transfer-Based Machine Translation (TB-MT) task.

1. Introduction

The effectiveness of fine grained grammatical distinctions at morpho-syntactic and semantic level is especially relevant cross-linguistically: the discussion of the aligned, bilingual Treebank, Italian (ITA) - Italian Sign Language (LIS) presented in these pages aims to provide linguistically motivated answers to two main questions:

1. are standard tagsets (e.g. Eagles, Monachini 1995, Tamburini 2007) sufficiently rich to account for (quasi-)deterministic rearrangement of constituents in a Transfer-Based Machine Translation (TB-MT) task?
2. is the phrase structure predicted by current linguistic frameworks (e.g. Minimalism, Chomsky 1995-2005, and Cartography, Belletti 2004, Cinque 2002 and Rizzi 2004) coherent with pervasive corpus-attested syntactic constructions and suitable for massive transformations between two fairly different languages?

In the first part of this paper (§2) we present the structure and the format we adopted to annotate the treebank, briefly discussing the set of features we used to code functional (e.g. topic, focus) and non-manual aspects (e.g. facial expression, movement velocity); in the second part of this paper we justify some radical linguistic assumption (e.g. head-marked, mainly flat tree-structures) on the basis of recent advances of Minimalist and Cartographic approaches. We tried to implement a version of tree structure that productively suits, as efficiently as possible, a TB-MT task: this means that the translation process is based on a structural reordering/pruning

procedure, driven by the leading idea that nothing in the structure must neither be created nor destroyed, but simply rearranged or scattered/collapsed (§3.3). This approach does not guarantee always an High Quality MT, but it results in fairly acceptable translations and it presents appealing computational advantages (§4).

2 The bilingual Treebank ITA-LIS

Despite the difficulty in defining a standard for coding a full transcription of a signed language[1], in building the ITA-LIS Treebank, we faced the problem of accounting, in a compact and meaningful way, for a complex parametric setup in order to exploit the annotated data from a Principles and Parameters (Chomsky 1981) point of view:

ITA	*LIS*
Head initial	Head final
Verb raising	No verb raising
Obligatory wh-movement	No wh-movement
Poor relative/PPs extraposition	Rich (obligatory?) relative/PPs extraposition
Rich clitic system	No clitics
No classifiers	Rich classifier system
Gender/number agreement	Spatial agreement

Table 1. Macro-parametric differences between ITA and LIS

Despite these differences there are also similarities, for instance, they are both *pro-drop* languages, they seem to allow (at least superficially) for a certain degree of variability in word order, they both show some rightward Heavy NP-shifting preferences. These parametric settings require a rich collection of empty elements (e.g. null subjects, traces, ellipses) and a consistent/computable solution to indicate referents and dependencies without losing any relevant linguistic information (e.g. (hanging-)topic/focus, argument doubling etc. Belletti 2004).

2.1 Corpus composition

The first release of the corpus is composed by 1018 Italian sentences extracted from public broadcast television news and translated/glossed to Italian Sign Language. 27 editions have been transcribed: 18 special editions written on purpose for LIS translation (shorter sentences, less complex structures[2]) plus 9 standard afternoon editions (standard Italian, without any special attention to the translation task). The ITA section of the Treebank has 17122 tokens (5391 distinct lexical items), while the LIS section counts 11056 tokens (3400 distinct lexical items). The asymmetry is due to the absence of various functional elements (e.g. articles, prepositions, auxiliaries etc.) as distinct lemmas in LIS (these elements are all coded by suprasegmental features, e.g. facial expressions) and to arguments/modifiers incorporation (e.g. the ITA equivalent of "to put a book on the shelf" is translated with the LIS equivalent of "to shelve a book"). The corpus is annotated using XML (Mana and Corazzari 2002), which ensures portability and permits a standard, flexible and human readable multilevel annotation. Structures can however readily (and univocally) be converted

[1] See Bergman et al. (2001).

[2] Roughly speaking, standard arguments order ((S)VO) is fairly maintained, words used are often high frequent lemmas, no relatives are employed and, in general, minimal NP modification is used (always locally); no parenthetical or long/run on sentences object are present.

into PENN (constituency) and TUT (dependency) format (Marcus et al. 1993, Bosco et al. 2000) to the detriment of some relevant linguistic distinction.

2.2 Signs, words and features

At the morphosyntactic level any single token (PoS) is enclosed under the tag <word> as follows:

(1) *<word id ="6" cat="V.ind.pres" lemma="essere" agree="3.s" role="head" subcat="copula"> è </word>*
(token: *è=is*, lemma: *essere=be*)

According to the Document Type Definition (DTD) we adopted (Appendix A), this is the list of attributes that can be specified for the tag <word> and their potential value:

(2) *id* it is an (unique) identity number for the node (nodes are recursively numbered in each tree from top to bottom, left to right);

ref traces, ellipses, pronominal elements and co-referent nodes in general have this attribute filled with the *id* of the reference node (it can be a relative specification: e.g. S-1 means the previous sentence, H-1 means the previous head, according to the numbering scheme just mentioned);

cat it is the classical PoS tag. Main tags are *N*(ouns), *V*(ebrs), *A*(djectivals), *ADV*(erbials), *D*(eterminers), *Q*(uantifiers). Each of them can be further subcategorized according to cartographic features (Appendix B): e.g. *V.ind.pres* classifies a *verbal* element, in *indicative* (vs. *subjunctive* vs. *infinitive* etc.) modality, at *present* (vs. *past* vs. *future* etc.) tense; *N.comm.count.inanim* classifies a *common* (vs. *proper*) nominal element, *countable* (vs. *mass*), *inanimate* (vs. *animate* vs. *animate-person*);

subcat it expresses obligatory thematic dependencies: *intrans*(itive) requires a subject grammatical position to be filled with an argument associated to the *agent* theta-role; *unacc*(usative) requires the grammatical subject position to be filled with a *patient* theta-role; *trans*(itive) requires two arguments positions subject and object to be filled respectively with an *agent* and a *patient* theta-role etc. (Appendix B); according to the Uniform Theta-role Assignment Hypothesis (Barker 1997) we do not need any further feature to identify univocally any dependency at the argumental level;

lemma it is the dictionary form of the token;

role three main dependency roles are allowed: (constituent) *head*, *arg*(ument) or *adj*(unct); *arg* can be *subj*(ect), *obj*(ect) *pred*(icative object), *ind*(irect object) (as in Bosco et al. 2000); *adj*s (and other dependencies) are listed in Appendix C;

lp left-peripheral (in the sense of Rizzi 1997) features such as *topic* and *focus* and other "edge" phenomena (Chomsky 2005) such as *expletive, doubling, extraposition* etc.;

expr (only in LIS) it expresses supra-segmental features such as eyebrows position (*eye-up*, *eye-down*), intensity of the sign (*slow*, *fast*, *minimize*, *exaggerate*) and the classifier system (*keep-support-hand*, *gaze-to-sign*, *cl.shape*, *cl.move*, *cl.position* etc. this is when a sign is not signed as reported in the dictionary but it "agrees" in shape, position etc. with another sign, Appendix D);

agree (only in ITA) person/gender/number features (e.g. *3.m.s* means third person, masculine, singular); (only in LIS) position features, organized by relevant spatial location such as eyes, mouth, chest etc. (e.g. *body_contact.mouth.left* means that the token is signed touching the mouth on the left);

sem it specifies a reference to the related MultiWordnet sense/synset (Bentivogli et al. 2002; if nothing is specified, the first synset associated to the lemma is picked out; this feature is still under implementation).

Id (used for co-reference and alignment), *cat* and *lemma* are obligatory, all the other attributes are optional.

Non-terminal nodes are coded with the tag <node> and they share the same attributes of the tag <word>. For mnemonic (and backward-compatibility issues) we used three standard categories to fulfill the "cat" attribute in <node>: NP (Nominal Phrase), VP (Verbal Phrase), AP (Adjectival/Adverbial Phrase). Such a simplification is not innocent, but it seems to be empirically and computationally tenable (Chesi 2007).

This is a sample of a tagged sentence:

(3) più difficile la situazione in Senato domani
more difficult the situation in Senate tomorrow
"tomorrow the situation in Senate will be more difficult"

```
<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>
```

And this is the LIS translation of the very same sentence:

(4) domani camera-Senato situazione difficile più
tomorrow room-Senate situation difficult more

```
<node cat="VP" id="2008-01-23.3" role="head">
    <word cat="ADV.time" id="11" lemma="domani" role="head">domani</word>
    <node cat="NP" id="8" role="adj.loc">
        <word cat="NE.org" id="10" lemma="senato" role="head">camera senato</word>
    </node>
    <node cat="NP" id="5" role="head">
        <word cat="N.comm.count.inanim" id="7" lemma="situazione" role="head">situazione</word>
    </node>
    <node cat="AP" id="1" role="adj.manner">
        <word cat="A.qualif" id="3" lemma="difficile">difficile</word>
        <word cat="ADV.limit" id="2" lemma="più">più</word>
    </node>
    <word agree="3.s" cat="V.ind.fut" id="4" lemma="essere" role="head" subcat="copula"/>
</node>
```

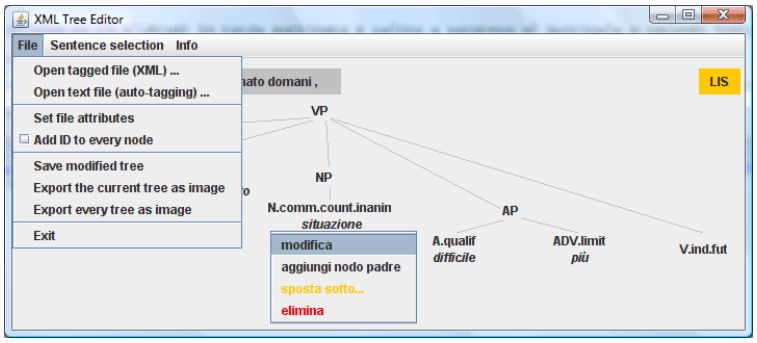
Notice that, despite their equivalence with the corresponding Italian words, characters within the tag <words> are simply indices that points, univocally, to a dictionary entry (that is, in fact, not simply a gloss, but a set of instruction to move an avatar, Bartolini et al. 2006); their relation with the corresponding Italian word is expressed only by the *id* field. In the case of homographs, *sem* is used to retrieve the correct item from the bilingual lexicon.

2.3 The annotation procedure

The morphosyntactic annotation consists of assigning to every token a <word> tag with the above mentioned features fully specified; as in other constituent-based annotations, words are grouped under the tags <node> to identify phrases. A well formed tree has one single node at the top. Every <node> has to bear a "cat" and a "role" specification and every well-formed node must be headed, which means it has one, and only one, <node> or <word> child with "role" equals to "head" (this can be phonologically null as the copula in (3)).

The XML structure is manipulated using a Java tool (XML Tree Editor[3]):

(5)



This tool can operate getting in input a text file (one sentence per line): it assigns to every sentence, automatically, a potential structure using a minimalist parser (based on Chesi 2007). Every structure that is automatically created can be graphically edited (nodes can be created, deleted, moved, replaced and every attribute can be modified). To guarantee consistency and reliability during the Treebank building, the grammar used by the parser is enhanced by rooted/terminal and auxiliary trees (as in Tree Adjoining Grammars (TAGs), Frank and Kroch 1995), i.e. previously tagged portions of sentences are ordered by frequency and used to help the parser retrieving the most likely structures.

3. Linguistic considerations

Evaluation of computational linguistic resources for Italian (EVALITA 2007) recently proposed a gold standard for the Italian PoS tagset (Tamburini 2007), and for the Constituency/Dependency classes/relations (Bosco et al. 2000) creating a lowest common denominator that includes widely used morphosyntactic/functional classes (e.g. PENN tag set). These standards are sufficiently rich and flexible to account for a wide range of linguistic phenomena, but not for a (quasi-)deterministic MT task between two languages parametrically as different as ITA and LIS (Table 1). The goal of this section is to highlight the major linguistic/computational necessities that induced a refinement (as minimal as possible) of such standards.

[3] The tool is freely available at http://www.ciscl.unisi.it/ricerca.htm

3.1 Bare Phrase Structure (BPS)

Within the Minimalist framework (Chomsky 1995-2005), many linguists assumed that lexical elements directly create constituency relations without projecting any non-terminal category (*Bare Phrase Structure* hypothesis, *Inclusiveness Condition*, Chomsky 1995); such a grammatical intuition (henceforth BPS) has been shown to be sound (Stabler 1997) and parsable (Harkema 1997) and it would dispense our grammar from using non-terminal (constituency) tags (e.g. NP, PP etc.) at all:

(6)

PENN-like	*BPS* (classic)	*BPS* (adopting Abney 1987)
(NP (ART the (N dog))	(dog (the) (dog))	(the (the) (dog))

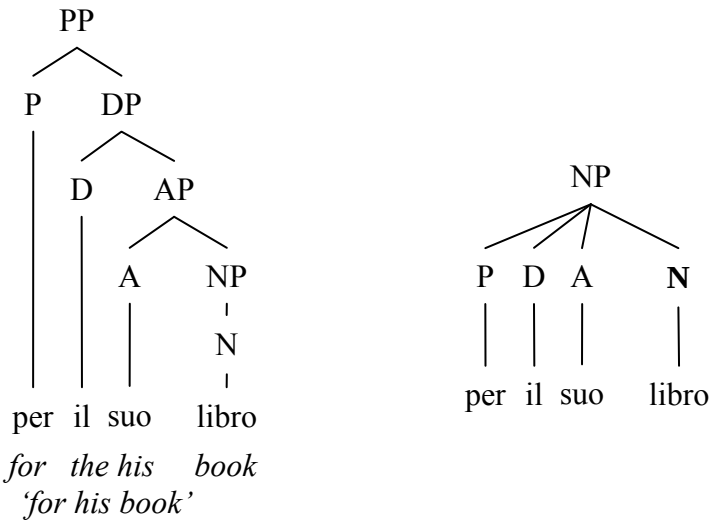
The selecting head (the noun in the classical X-bar generative theory, the determiner, after Abney's influential proposal) projects over the selected element. In this case, we would expect any lexical item to be marked for selection within the lexicon. Since the selecting element is always a head (by definition) we guarantee that the projecting node is the head of the phrase (i.e. an NP node is in fact the projection of a N head, (6).BPS-classic). Notice that while in standard minimalist approaches the projection system results in a binary operation (i.e. *merge*) that strictly produce binary branching trees, we assume that the constituents can have more than one sister. In the following paragraph we will defend the idea that even if we do not assume a binary branching constraint (Kayne 1983), binary branching structures can be readily retrieved from the proposed tree and hence, BPS-related assumptions can be kept. On the other hand, as introduced in §2.2, the fact that we mark nodes as VPs, NPs or APs (*cat* feature in our xml structure), is not against the BPS idea since we can unambiguously track the projecting heads node by node (this is so because every node has exactly one single head).

3.2 Cartography of functional projections

From an empirical point of view flat trees have often been challenged in literature since non-predictive with respect to many relevant phenomena (e.g. coordination and gapping, binding etc. Kayne 1983); on the other hand, having flat structures reduces the ambiguity in the lexicon[4] and allows us to retrieve, with a minimal search, every relevant feature in a given phrase (Adger 2007, Chesi 2007). This tension seems to be solvable if we accept the cartographic hypothesis (e.g. Cinque 2002): order and hierarchy are in fact tightly related and universally constrained; superficial "free" word order is the epiphenomenon of a sequence of movements that target functional/peripheral (Rizzi 1997) positions. The attribute *lp* (§2.2) exactly expresses these "extra" features and prevents us from implementing a full projection of every functional node, including their potential landing site in the left periphery: about forty distinct positions in the functional VP domain (Cinque 2002) can be collected under the same node (i.e. these forty nodes are optionally present and, when present, all dominated by the same VP node, without requiring any selection mechanism within the lexicon) keeping their cartographic (sub)category (e.g. for the adverbial domain: ADV.manner, ADV.temp, ADV.neg, ADV.asp etc.). The example below shows the tree-translation between standard approaches and the one implemented within the Siena University Treebank (SUT).

[4] This is because of the selection mechanism proposed by Chomsky and formalized by Stabler: if each node has to be marked for selecting its sister category, having or not having an optional adjective, for instance, between the determiner and the noun, would duplicate the number of determiners: one that selects the adjective and another one that does not (Chesi 2007).

(7) Standard tree SUT simplified tree



The subcategorization of main functional categories (e.g. adverbs) directly expresses the dependency (i.e. the xml attribute "role") of the <word> element within the phrase with respect to the phrase head (otherwise, within the <node> tag, the "role" attribute is specified as a specific "adjunct" category; this is, for instance, the case of adverbial PPs, that, following our guidelines, are simply tagged as NPs). Building an extensive treebank with such information could then turn out to be a precious tool also for evaluating quantitatively the predictions of the cartographic approach.

4. Evaluation of the Treebank from a TB-MT perspective

The goal of this paper was to present in a fairly intuitive and compact way the process of treebank building and the theoretical assumptions that justified certain choices. In this final section we evaluate in which sense the standard we proposed is different from the alternative Eagles/EVALITA tagset (Monachini 1995, Tamburini 2007) and TUT set of dependencies (Bosco et al. 2000) (§4.1). Then we will verify if such refinements are productive when we try to extract alignment rules from the treebank that should be suitable for a transfer-based MT task (§4.2).

4.1 (Minimally) different standards

Despite main categories such as Verbs, (Pro)Nouns, Articles, Prepositions, Adjectives, Adverbs are consistently adopted following the standard discussed in Monachini (1995) and Tamburini (2007), few differences at sub-categorial and functional level are worth to be reported and justified: as for the functional level, for instance, articles are collected under the PoS D(eterminer) together with quantifiers and demonstratives (see Appendix B for a full list) in order to capture some cartographic intuition ("determiner" vs. "adjectival" field); (subordinating) conjunctions as well as prepositional subordinators are included under the PoS C(omplementizer) again to comply with cartographic ideas ("left-peripheral" Vs. "inflectional" field). On the sub-categorization side, the table below highlights some substantial expansion of the proper name and adjectival classes (again, refer to Appendix B for the whole picture):

Category	*Eagles*	*SUT*
Proper Names	SP@NN	N.prop.anim.person.last/first N.prop.inanim.city ...
Adjectival forms	A (adjective), AP (possessive adjective)	Adj.deict, Adj.dem, Adj.nation, Adj.num.ord, Adj.num.card, Adj.poss, Adj.qualif …

These distinctions are mainly justified by the task we are dealing with: (quasi-)deterministic rearrangement of constituents in a TB-MT task; these distinctions are in fact crucial since names of persons or cities have to be prefixed in LIS by the correct classifier, "person" or "city" respectively. On the other hand, subcategorizing adjectival forms gives us the opportunity to reorder correctly (in standard contexts) these elements in LIS:

(8) Adj.num < Noun (head) < Adj.poss < Adj.dem/deict < Adj.nation < Adj.qualif

On the dependency side, a differently structured set of relations (according to the categories of *head*, *arguments* and *adjuncts*) allows us to correctly predict phenomena such as relative extrapositions or PP clefting in LIS which would be less transparent under the distinction *functional arguments* (e.g. locatives are considered arguments under the label of "indirect complements" much as the subject and the direct object in Bosco et al. 2000).

The necessity for such distinctions becomes clear analyzing the head directionality parameter (table 1, §2): reordering is massive between ITA and LIS and the linguistic assumptions we discussed allows us to deal in a computationally elegant way with this, since relevant constituents are readily accessible within just one level of inspection. This allows us, for instance, to capture in-corporation (9), ex-corporation of arguments/adjuncts (10) analyzing only immediate constituents within a single phrase (θ expresses the thematic requirements of the head; $i\theta$, indicates an internal, lexical, satisfaction of such requirement):

(9) (ITA) $[_{VP}\ [_{head}\ \theta_1\ ]\ [_{arg.obj}]]$ → $[_{VP}\ [_{head}\ i\theta_1]]$ (LIS)
$[_{VP}$ mettere $[_{arg.obj}$ una firma$]]$ → $[_{VP}\ [_{head}$ firmare$]]$
put a sign sign

(10) (ITA) $[_{VP}\ [_{head}\ i\theta_1]\ ]$ → $[_{VP}\ [_{arg.obj}]\ [_{head}\ \theta_1]]$ (LIS)
$[_{VP}$ dimettere] → $[_{VP}\ [_{arg.obj}$ carica$]\ [_{head}$ rinunciare$]]$
dismiss position leave

Standard argument/adjuncts reordering (11) as well can be readily decided locally without inspecting further constituents:

(11) (ITA) $[_{VP}\ [_{arg.subj}]\ [_{V\text{-}head}]\ [_{arg.obj}][_{adj.temp}]]$ →
(LIS) $[_{VP}\ [_{adj.temp}]\ [_{arg.subj}]\ [_{arg.obj}]\ [_{V\text{-}head}]]$

Then a richer (cartographic) subcategorization allow us to extract from the corpus non-ambiguous reordering rules of adjuncts:

(12) (ITA) $[_{VP}\ [_{V\text{-}head}]\ [_{adj.manner}]\ [_{adj.matter}]]$ → $[_{VP}\ [_{adj.manner}]\ [_{V\text{-}head}]\ [_{adj.matter}]]$ (LIS)

It should be clear then that having more fine grained categories and features allows us to extract more specific transfer-based rules. There is however a drawback in freely multiplying features and categories: the data required to extract statistically reliable information would grow considerably. We attempted to solve this problem using an

hierarchical approach to categories/features expansion: information on the distribution of upper level categories are much more easily retrievable, then using this data guarantees a fairly robust TB-MT approach. On the other hand, accuracy is pursued rewarding sub-categorial distinctions, with respect to main categories, whenever they reach a reasonable frequency threshold.

4.2 Quality of translation

To give an example of the real input-output we should expect from a TB-MT system as the one we built using our aligned corpus, we report below an example that shows a simple case:

(13) Input string:
il presidente del consiglio parla con un segretario
the president of the council talks with a secretary

Output of the parsing analysis[5]:

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<node agree="3.s" cat="VP" role="head">
  <node agree="m.s" cat="NE.per" role="arg.subj">
    <word agree="m.s" cat="D.art.def" id="2" lemma="il">Il</word>
    <word agree="m.s" cat="N.comm.count.person" id="3" lemma="presidente" role="head">Presidente</word>
    <node agree="m.s" cat="NP" id="4" role="adj.matter">
      <word agree="m.s" cat="P.matter" id="5" lemma="di+il">del</word>
      <word agree="m.s" cat="NE.org" id="6" lemma="consiglio" role="head">Consiglio</word>
    </node>
  </node>
  <word agree="3.s" cat="V.ind.pres" lemma="parlare" role="head" subcat="transitive">parla</word>
  <node agree="m.s" cat="NP" role="adj.comitat">
    <word cat="P.comitat" lemma="con">con</word>
    <word agree="m.s" cat="D.art.indef" lemma="un">un</word>
    <word agree="m.s" cat="N.comm.count.anim" lemma="segretario" role="head">segretario</word>
  </node>
</node>
```

Output of the tranfer-based MT process:

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<node cat="VP" role="head">
  <node cat="NE.per" id="0" role="arg.subj">
    <word cat="N.comm.count.person" id="2" lemma="presidente" role="head">persona presidente</word>
    <node cat="NP" id="3" role="adj.matter">
      <word cat="NE.org" id="5" lemma="consiglio" role="head">Consiglio</word>
    </node>
    <word cat="END.comma" id="7" lemma=",">,</word>
  </node>
  <node cat="NP" role=" adj.comitat">
    <word cat="N.comm.count.anim" lemma="segretario" role="head">persona segreteria</word>
  </node>
  <word cat="V.ind.pres" lemma="parlare" role="head" subcat="transitive">dire</word>
</node>
```

Despite this very simple example, such approach allows us to deal with rather subtle phenomena: for instance, extraposition and leftward-movement are constrained in a

[5] NE.per are Personal Named Entities, NE.org are Organization Named Entities.

very productive way by flattening the structure: assuming that the attachment point of the extraposed relative/PP is the immediate upper phrase (14), we can capture 96% of extraposed constituents; this is true also for genitive constructions (15):

(14) (ITA) $[_{VP}\ [_{arg.subj}[_{adj.rel.restr}]]\ [_{V\text{-}head}]] \rightarrow$
(LIS) $[_{VP}\ [_{arg.subj}]\ [_{V\text{-}head}]\ [_{adj.rel.restr}]$

(15) (ITA) $[_{NP}\ [_{N\text{-}head}]\ [_{arg.subj}[_{N\text{-}head}]]] \rightarrow [_{NP}\ [_{N\text{-}head}]_i\ [_{N\text{-}head}]\ [_{arg.subj}[_i]]]$ (LIS)
la foto di Gianni — Gianni foto sua
the picture of John — *John picture his*

Moreover, using empty elements (e.g. null-subjects, reduced relatives etc.) and a (relative) referential mechanism allow us to extract rules for re-integrating the referents also in discontinuous dependents:

(16) [Il rappresentante$_i$ [di profumi] [che$_i$ è venuto ieri]] $\rightarrow$
The perfume salesman *that came yesterday*

$[_{NP}\ [_{N\text{-}head}]\ \ldots\ [_{NP/RC}\ \text{Relative\_Pro}_{head\_of\_the\_dominating\_NP}\ \ldots]]$

In the end we attempted to make a human evaluation of the TB-MT system: a set of 50 sentences (Appendix E) which the system has not been trained on, has been semi-automatically analyzed and then automatically translated according to the rule extracted from the aligned corpus. A native speaker evaluated the provided translations with respect to word order soundness[6], on a scale from 0 to 3 (0=incomprehensible, 1=comprehensible but sub-standard, 2=comprehensible, 3=good). The translations received an average score of 1.58, which is not a bad result at all for a naïf TB-MT system.

4.3 Concluding remarks

In the end, we showed that the proposed structures/categories, inspired by main current generative frameworks (Minimalism, Chomsky 1995-2005, and Cartography, Belletti 2004, Cinque 2002 and Rizzi 2004) can be coherently implemented in a bilingual aligned treebank ITA-LIS. The TB-MT task seems to take advantage of such a rich structure and the translation provided seems to be fairly acceptable by native speakers. Obviously more tests are required on the word sense disambiguation side and the treebank should be significantly augmented from a quantitative point of view. These first results however seem to show that the undertaken mission is fully promising.

[6] Since some of the lexical items were not present neither in the corpus nor in the aligned bilingual lexicon, we could not expect the system to make the correct lexical choice in these cases.

Appendix A – XML DTD

The Document Type Definition (Siena University Treebank, Version 1.0) is defined as follows (the DTD filename referred by the XML files in the treebank is "SUT.dtd"):

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT text (node|expression)+>
<!ELEMENT node (node|word)*>
<!ELEMENT expression (#PCDATA)>
<!ELEMENT word (#PCDATA)>
<!ATTLIST expression id CDATA #REQUIRED>

<!ATTLIST text id CDATA #REQUIRED>
<!ATTLIST text lang CDATA #REQUIRED>
<!ATTLIST text type CDATA #REQUIRED>

<!ATTLIST node id CDATA #IMPLIED>
<!ATTLIST node cat CDATA #REQUIRED>
<!ATTLIST node subcat CDATA #IMPLIED>
<!ATTLIST node ref CDATA #IMPLIED>
<!ATTLIST node role CDATA #IMPLIED>
<!ATTLIST node agree CDATA #IMPLIED>
<!ATTLIST node expr CDATA #IMPLIED>
<!ATTLIST node lp CDATA #IMPLIED>

<!ATTLIST word id CDATA #IMPLIED>
<!ATTLIST word cat CDATA #REQUIRED>
<!ATTLIST word subcat CDATA #IMPLIED>
<!ATTLIST word ref CDATA #IMPLIED>
<!ATTLIST word agree CDATA #IMPLIED>
<!ATTLIST word role CDATA #IMPLIED>
<!ATTLIST word lemma CDATA #IMPLIED>
<!ATTLIST word expr CDATA #IMPLIED>
<!ATTLIST word lp CDATA #IMPLIED>
<!ATTLIST word sem CDATA #IMPLIED>
```

The following attributes are only used by the XMLTreeViewer tool and are never displayed in the user-accessible XML structure:

```
<!ATTLIST node tmpid CDATA #IMPLIED>
<!ATTLIST node x CDATA #IMPLIED>
<!ATTLIST node y CDATA #IMPLIED>
<!ATTLIST word tmpid CDATA #IMPLIED>
<!ATTLIST word x CDATA #IMPLIED>
<!ATTLIST word y CDATA #IMPLIED>
```

Appendix B – Attribute-Value constraints

This is the list of the main attributes (linguistic features) and their possible values (SUT Version 1.0). The number that precedes the value indicates the absolute order of the features when they are concatenated under the same attribute (concatenation of features is not a linguistically motivated solution, it simply solves a backward compatibility issue; most of the time every row would deserve an independent attribute specification, some other time concatenated features should be grouped in a different way (e.g.); the next version of the tools should consider this issue).

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

Sem and *Lemma* (as *Id* and *Ref*, §2.2) will be omitted from the following tables since the same values/constraints discussed here will apply.

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	as for Nouns	
Role	[adj]	adjunct

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

Appendix C - Functional Dependencies

The set of dependencies used to annotate the relation between phrases is the following one:

- 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?*]

Appendix D – Special features for tagging Sign Languages

Sign Languages require an enriched set of features to express properties that are not usually present in oral languages (e.g. morpho-syntactic Agreement in Sign Language is on a spatial dimension rather than on a gender dimension as in Oral Languages).
In the table below we report the set o features used to express agreement and other functional features (*lp* attribute in our xml files):

What	Feature	Brief Explanation
Agree		We decide to refer to the space agreement as a "3-dimensional" space in a non-conventional sense:
	body_contact	The first dimension is the contact with the body. By default a sign is not expressed touching a specific part of the body (unless explicitly marked in the lexicon);
	forehead/eyes-nose/mouth/neck/ chest/stomach	A second dimension is the height of the sign: by default a sign is expressed in the neutral space, that is in front of the chest; otherwise it can be signed at the forehead level or at the eyes-nose, mouth, neck, chest or stomach levels

	left/right	On a horizontal dimension a sign is expressed by default in the neutral space, that is right in front of the chest; otherwise we can specify a left or right position No agreement information means that the sign is expressed in the neutral space; otherwise non-default dimensions are concatenated e.g. *body_contact. mouth.left*
Classifiers	cl.shape/cl.space/ cl.movement	The classifier system indicates when a sign is not expressed as coded within the lexicon; shape, space and movement are the feature that the modified sign inherits from the dependent sign in the context (the head of the phrase if not explicitly marked)
Special functional features	eye-up	Eyebrows up (yes-no question)
	eye-down	Eyebrows down (wh-/rhetorical questions)
	keep-support-hand	keep-support-hand
	neg	head shaking expressing negation
	past/fut	movement to express past (toward the shoulder) and future (from the shoulder) times
	slow/fast	velocity modification of the sign (e.g. depending of the strength adverbial modifiers)
	exaggerate/minimize	exaggerate/minimize the movement of the sign (expresses diminutives, augmentatives features)
	now	gaze at the neutral space (it indicates the present time)
	gaze-to-sign	the gaze directed to the sign marks the familiarity with the object just signed
	labialization	indicates that a specific labialization is required (usually when a word is finger-spelt)

Appendix E – Machine Translation Test

The alignment rules extracted from the corpus have been used to train a MT system. This system has been exposed to 50 new parsed sentences and the pairs original Italian sentence - MT output in LIS has been evaluated by a native speaker. The score ranged from 0 (incomprehensible) to 3 (good) (1 = comprehensible but sub-standard; 2 = acceptable);
This is the raw result of the test: 3 - 22%; 2 - 26%; 1 - 40%; 0 - 12%; mean: 1.58

ID	*Original Italian Sentence*	*Output in LIS*[7]	*Score*
1	accade che qualche giovane si accontenti	pro giovane accontentare accadere	**3**
2	cinquecento studenti sono ospitati in collegi universitari	collegio;PL universitario cinquecento persona studiare ospitare fatto	**3**
3	una città che non riesce ad accogliere i giovani è destinata a morire	pro accogliere persona giovane riuscire no città morire destinato	**3**
4	ha deciso , infatti , di non applicare il decreto	proprio decreto applicare no decidere fatto	**3**
5	così , è stato deciso di battere il bosco	bosco battere decidere fatto	**3**
6	i commissari di gara lo hanno accusato di avere intralciato Massa	gara commissario;PL accusare Massa intralciare fatto	**3**
7	quando il vaccino terapeutico risulterà disponibile	data vaccino curare c'è risultare	**3**
8	riscaldarsi quest' inverno sarà veramente costoso	inverno riscaldare+si costoso veramente	**3**
9	potrà anche estinguere il mutuo	potere anche mutuo estinguere	**3**
10	la squadra di Detroit ha quasi sempre vinto la partita	squadra Detroit partita vincere quasi	**3**
11	io ci ho sempre provato	io provare fatto	**3**
12	come fate a offrire residenze a prezzo calmierato ?	prezzo calmierato residenza;PL offrire fare	**2**
13	gli affitti continuano ad essere cari	affitto;PL continuare caro	**2**
14	gli incidenti sono tutti da attribuire al fattore umano	tutti incidente attribuire fattore umano	**2**
15	è stata subito sottoposta a terapia malarica	subito terapia malarico fatto	**2**
16	la pace era già stata raggiunta da Heider	già problema Heider raggiungere fatto	**2**
17	lui ha deliberatamente frenato troppo tardi	troppo tardi pro frenare fatto deliberatamente	**2**
18	la fede dipende da Dio , e da Dio solo	fede dipendere Dio + Dio solo	**2**
19	si tratta di saper distinguere le emozioni	emozioni;PL sapere distinguere	**2**
20	siamo già in recessione	già recessione	**2**
21	i fondi per l' Africa si sono drasticamente ridotti	fondo;PL motivo africa ridurre fatto drasticamente	**2**
22	lo si era capito già in partenza	già partenza capire fatto	**2**
23	si chiamano nuovi acquisti perché devono portare qualcosa di nuovo	acquisto nuovo chiamare motivo dovere qualcosa nuovo portare	**2**
24	il contesto sociale in cui si è nati e cresciuti	nato fatto + crescere fatto situazione sociale	**2**
25	ha deciso di non applicare il decreto	decreto applicare no decidere fatto	**1**
26	i manager che falliscono saranno messi da parte	pro fallire manager;PL mettere fatto	**1**

[7] *pro* indicates a deictic sign to a position in the space where the referred object has been previously signed. *;PL* indicates that the sign that precedes it has to be repeated (according to its plural status). + indicates the sign used for the conjunction of two expression. All suprasegmental features discussed are not included in the simple text transcription.

27	si è così arrivati all' individuazione di numerosi immobili	arrivare fatto immobile;PL numeroso individuazione	**1**
28	troverà ad attenderlo una lunga fila di bandiere italiane	bandiera;PL italia fila lungo trovare attendere	**1**
29	potrebbe essere sciolta la prognosi sulla sopravvivenza	potere sopravvivenza prognosi sciogliere fatto	**1**
30	sembra che ci fossero anche pietre difficili da individuare	pro anche pietra;PL individuare difficile sembrare	**1**
31	quando ieri gli ha annunciato che voleva parlare con lui , è rimasto in silenzio	data ieri pro volere parlare pro silenzio annunciare fatto	**1**
32	una parte della Curia fiorentina si accorgerà che avevamo ragione	curia fiorentino pro ragione avere accorgere	**1**
33	per stabilire dove stia la ragione e dove stia il torto	motivo ragione stare + torto stare stabilire	**1**
34	è stato Hamilton a sbagliare	sbagliare Hamilton	**1**
35	ammetto di aver sbagliato al via	sbagliare fatto ammettere	**1**
36	tutti voi siete una sola persona in Cristo	tutti voi Cristo persona sola	**1**
37	il tempo deve diventare la misura della vostra pazienza	dovere periodo pazienza vostro misura diventare	**1**
38	si consiglia di limitare il consumo di queste verdure pronte	consumo verdura;PL pronto limitare consigliare	**1**
39	un canarino è evidentemente il migliore rimedio contro le preoccupazioni atomiche	canarino contro preoccupazione;PL atomico rimedio buono evidentemente	**1**
40	ne abbiamo già parlato anche troppo	già ne parlare fatto anche troppo	**1**
41	non ho mai pensato di segnare in quel modo	segnare pi modo pensare fatto no contro	**1**
42	non ha cercato di segnare con la mano	mano segnare cercare fatto no	**1**
43	ora ci chiediamo se sia giusto questo turn-over massiccio	ora se turn-over massiccio giusto chiedere	**1**
44	a un gruppo di scrittori emiliani viene assegnato un prodotto tipico	prodotto tipico scrittore;PL emiliano gruppo assegnare fatto	**1**
45	il locatore chiede all' inquilino di versare ulteriori somme	locatore versare somma;PL ulteriore inquilino chiedere	**0**
46	gli affitti che gli studenti si trovano a dover pagare	affitto;PL	**0**
47	è seguito da un tutor che lo aiuta a orientarsi nella scelta dei corsi	pro orientare corso;PL scelta aiutare tutor;PL	**0**
48	imperversano anche le locazioni in nero	anche nero locazione;PL imperversare	**0**
49	diventa sempre più difficile venire a studiare nel capoluogo lombardo	venire capoluogo persona lombardo studiare difficile più diventare	**0**
50	è effettivamente un' azione comune quella che proponiamo	pro proporre diventare azione effettivamente	**0**

References

Abney, S. 1987. *The English noun phrase in its nominal aspect*. Doctoral dissertation, MIT Press.

Adger, D. 2007. *A minimalist theory of feature structure*. http://ling.auf.net/lingBuzz/000583

Baker, M. 1997. "Thematic roles and syntactic structure". In *Elements of grammar: Handbook in generative syntax*, ed. Haegeman, L. Dordrecht: Kluwer: 73-137.

Bartolini S., Bennati P., Giorgi R. 2006. *Bluesign-2, il nuovo visualizzatore portatile per la Lingua Italiana dei Segni*, in Proceedings of "51th Corso Nazionale di Studio, Formazione e Aggiornamento dell'AIES". Siena:Cantagalli:140-145.

Belletti, A. 2004. *Structures and Beyond.* Oxford, Oxford University Press.

Bentivogli, L., E. Pianta & C. Girardi, 2002. Multiwordnet: developing an aligned multilingual database. *First International Conference on Global WordNet, Mysore, India*.

Bergman, B., P. Boyes-Braem, T. Hanke & E. Pizzuto 2001. *Sign Transcription and Database Storage of Sign Information*. Special issue of Sign Language & Linguistics, 4:1/2.

Bosco, C., V. Lombardo, D. Vassallo, & L. Lesmo, 2000. "Building a treebank for Italian: a data-driven annotation schema." *Proceedings of the Second International Conference on Language Resources and Evaluation LREC,* 99-106.

Chesi, C. 2007. "An introduction to Phase-based Minimalist Grammars: why move is Top-Down from Left-to-Right". *STiL - Studies in Linguistics*, 1.

Chomsky, N. 1981. "Principles and parameters in syntactic theory." *Explanation in Linguistics: The Logical Problem of Language Acquisition,* 32-75.

Chomsky, N. 1995. *The Minimalist Program*. MIT Press.

Chomsky, N. 2005. On Phases. *Manuscript, MIT*.

Cinque, G. 2002. *Complement and adverbial PPs: implications for clause structure*. Abstract, University of Venice.

Cinque, G. 2002. *The cartography of syntactic structures. Vol. 1, Functional structure in DP and IP,* Oxford University Press.

Frank, R. & A. Kroch. 1995. "Generalized transformations and the theory of grammar." *Studia Linguistica* 49:103-151.

Harkema, H. 2001. "Parsing Minimalist Languages." University of California Los Angeles.

Kayne, R. S. 1983. Connect & M. Calcagno. 2001. Parasitic gaps in English: some overlooked cases and their theoretical implications. In *Parasitic Gaps*, ed. P. Culicover and P. Postal, 181-222. Cambridge, Mass.: MIT Press.

Kayne, R. S. 1983. *Connectedness and binary branching*. Foris Publications.

Mana, N. & O. Corazzari, 2002. "The lexico-semantic annotation of an Italian Treebank." *Proceedings of LREC 2002*.

Marcus, M. P., M. A. Marcinkiewicz & B. Santorini, (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*, 19(2), 313-330.

Monachini, M. 1995. *ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines*. Technical report, Pisa.

Rizzi, L. 1997. The fine structure of the left-periphery. In *Elements of Grammar: Handbook in generative syntax*, ed. Haegeman, L. Dordrecht: Kluwer. 1997, 281-337.

Rizzi L. ed. 2004. *The Cartography of Syntactic Structures,* Oxford University Press.

Stabler, E. 1997. "Derivational minimalism." *LECTURE NOTES IN COMPUTER SCIENCE* 68-95.

Tamburini, F. 2007. *Evalita 2007: The Part-of-Speech Tagging Task*. Proceedings of EVALITA 2007.