VIT – Venice Italian Treebank: Syntactic and Quantitative

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli
University Ca' Foscari
Dept. Language Sciences - Laboratory Computational Linguistics

Abstract

In this paper we will describe VIT (Venice Italian Treebank), created at the University of Venice. We will focus on the syntactic-semantic features and on the quantitative analysis of the data of our treebank comparing them to other treebanks. In general, we will try to substantiate the claim that treebanking grammars or parsers is dramatically dependent on the chosen treebank; and eventually this process seems to be dependent either from substantial factors such as the adopted linguistic framework for structural description or, ultimately, the described language.

1 Introduction

In this paper we will describe VIT (Venice Italian Treebank), a treebank consisting of 320.000 words created by the Laboratory of Computational Linguistics of the Department of Language Sciences of the University of Venice. The VIT Corpus consists of 60.000 words of spoken text and of 270.000 words of written text. In this paper we will restrict our description to the characteristics of written texts of our Treebank. Syntactic annotation was accomplished through a sequence of semi-automatic operations followed by manual validation. The first version of the Treebank was created in the years 1985-88 – its rules were used to build a context-free parser for a speech synthesizer (Delmonte & dolci). The theoretical framework behind our syntactic representation was X-bar theory. Schematically speaking, X-bar theory (we refer here to the standard variety presented in LFG theory) prefigures an organization of the type head and head-projections where each head is provided with a bar in hierarchical order: in this way the node on which a head depends is numbered starting from 0 and the subsequent dominant nodes have a bar, two bars and if necessary other bars. The hierarchical organization of the theory consists of the following abstract rewrite rules:

1.1 Theoretical scheme of X-bar theory rules

CP --> Spec, Cbar
Cbar --> C0, XP
XP --> Spec, Xbar
Xbar --> X, Complements/Adjuncts
C0 --> Complementizer
X --> Verb, Adjective, Noun, Adverb

The first choice we operated had to do with the internal organization of the specifier of the NP that, in case of non-phrasal constituents, can consist of one or more linguistic elements belonging to different minor syntactic categories as reported below:

1.2 Atomic vs Structured Specifier

Spec--> Determiners, Quantifiers, Intensifiers

The choice to have a Spec structure was too difficult an option to pursue, so we decided to leave minor non-semantic constituents that stood before the head in an atomic form, unless it required a structure of its own, which could apply for quantifiers. Besides, semantic heads such as adjectives and adverbs have their own constituent structure. For sentence level the following X-bar-like scheme was employed:

1.3 X-bar rules for sentence level

CP --> Spec(dislocated constituents), Cbar
Cbar --> C0, S
S --> NP<SUBJect>, I0, Complements
C0 --> Complementizer
I0 --> Finite Tensed Verb

Tensed verb takes a separate structure we have called IBAR - or IR_INFL ("unreal" verb) when the verb is either in future, conditional or subjunctive form- and that can consist of more elements added to the constituency level of the tensed verb:

1.4 Verbal head structure

Ibar -->Verb – auxiliary verbs, modals, clitics, negatives, adverbials (also in a PP form)

Other specialized structures will be discussed further on, but now it is important to note that our representation does not employ a VP structure level: in fact, we preferred to analyse verbal group as directly positioned on the same level of S, where there will also be a NP-Subject, if syntactically expressed, and a complement structure subcategorized according to different types of complements. By doing this, VIT followed NEGRA, the German treebank, also in the sense of specializing major non-terminal constituents, as discussed in the sections below. While on the contrary PennTrebank (hence PT) differs for a less detailed and more skeletal choice, as specified in the PT guidelines.

Having a more specialized inventory of constituents was done also in view of facilitating further conversion into dependency structure. In particular, having a specialized node for tensed clauses, which is different from the one assigned to untensed ones, allows for better treatment of such constituent, which, as will be shown below, allows for some of its peculiar properties to be easily detected. Moreover, by assuming that tensed the verb compound – IBAR/IR_INFL - is the

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

sentence head, thus being is in line with a number of theoretical frameworks and allowing a much easier treatment in the dependency grammar scheme, where the head of the VP is also the head of S. Differently from what happens with PT, in VIT the verbal head doesn't have to be extracted from a substructure because it's already at S level: on the contrary, in PT the head could be the leaf of many different VP nodes depending on how many auxiliaries or modals precede the main lexical verb. In our case, for every further operation of transduction in structures and dependencies, the number of levels to keep under control is lower in comparison to the task of detecting the relations between a Head-root and Head-dependents.

Adding a VP node that encompasses the Verbal compound and its complement was not a difficult task to carry out. We have then produced an algorithm that enables the transformation of the entire VIT without a VP node into a version that conversely has it, but only in those cases where it is allowed by the grammar. In this way we successfully removed all those instances where the verbal group IBAR/IR_INFL is followed by linguistic material belonging to the S level, such as phrasal conjunctions, PP adjuncts or parenthetical structures. By doing this we were able to identify about 1000 clauses out of the total 16000 where the VP node hasn't been added.

In the second part of the paper, we will discuss the quantitative data concerning the written portion of VIT and the constituents present in the 10.200 utterances of its Treebank; the crude data are displayed in **Table 1.** below and will be compared to those in PT. Number of tokens indicated is the original one and does not take into account the number of additional tokens added because of split amalgams and cliticized verbs which amount to some 20k additional tokens.

In particular, we will focus on some structures that are interesting from a parsing point of view and are called "stylistic" structures. In the table below we also listed subtypes of sentence F – i.e. Sentence with Null Subject -, in order to highlight its internal distribution.

Constituent type	**Constituent Label**	**No. of occurences**
Total Utterances		**10.200**
Total of Tokens		**256.365**
Nominal Phrase	**SN**	**69.580**
Prepositional Phrase	**SP**	**21.393**
Prepositional Phrase with prepositions DI/DA	**SPD/SPDA**	**20.592**
Adjectival Phrase	**SA**	**21.205**
Adverbial Phrase	**SAVV**	**4571**
Quantified Phrase	**SQ**	**2523**
Comparative Phrase	**SC**	**520**
Verbal Group with Tensed Verb	**IBAR**	**13.404**
Verbal Group with Unreal Verb	**IR_INFL**	**2526**

Coordinate Structure for Constituents – heads with conjunction or punctuation	**COORD**	**5703**
Sentence	**F**	**15.851**
Subordinated Sentence with Subordinator	**FS**	**1063**
Coordinate Sentence with conjunction	**FC**	**3718**
Parenthetical, Apposition with Punctuation – Adjuncts Constituent	**FP**	**4381**
Interrogative Sentence with/without Interrogative Pronoun	**FINT**	**585**
Dislocated/preposed constituents, adjunct const.	**CP**	**4906**
Dislocated/preposed constituents, adjunct const.	**CP_INT**	**203**
Complement Sentence with/without Complementizer	**FAC**	**956**
Infinitival Clause/Participial Clause/Gerundive Clause	**SV2/SV3/SV5**	**7568**
Relative Clause with Relative Pronoun	**F2**	**3425**
Direct Speech with Punctuation – Any Constituent	**DIRSP**	**1101**
Sentence Fragment	**F3**	**3552**
Complement governed by Transitive Verb	**COMPT**	**11.478**
Complement governed by Intransitive Verb	**COMPIN**	**5580**
Complement governed by Copulative Verb	**COMPC**	**3886**
Complement governed by Passive Verb	**COMPPAS**	**340**
Aux-to-comp Constituents/Verbal Group with Tensed Auxiliary	**TOPF/AUXTOC**	**19**
Sentence with Null Subject	**F-[IBAR/ IR_INFL**	**9756**
Non-Phrasal Constituent Total		**183.301**
Phrasal Constituents Total		**47.328**
Constituents total		**230.629**

Table 1. Quantitative Data of VIT Constituents

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

2 A Quantitative Study of VIT

In a recent paper, Corazza et al. (2004) use a portion of VIT – 90.000 tokens produced in the National SI-TAL project, also called ISST – to verify the possibility to train a statistic-probabilistic parser on the basis of procedures already experimented in English with PT by Collins and Bikel. Since results obtained are preliminary and quite scarce (inferior to 70% accuracy), the authors wonder whether the poor performance might be due to intrinsic difficulties in the structure of the Italian language, to the different linguistic theory that has been adopted (cf. the lack of a VP node) or to the different tagset adopted, more detailed if compared to the one used in the PT.

According to what stated by Bikel regarding Collins' work, still a landmark for the creation of probabilistic parsers, the work done for the creation of a language model is to be anticipated by an important phase of preprocessing. This means that in order to produce the language model one does not work on the raw data of a treebank, but on a version modified on purpose. Collin's aim was to capture the biggest amount of regularities with the smallest number of parameters.

Probabilities are associated to lexicalized structural relations, i.e. structures where the head of the constituent to encode is present, with the aim of helping the parser to make decisions concerning the choice of arguments vs. adjuncts, of levels of attachment of a modifier and other similarly important matters otherwise difficult to capture when using only tags. For this purpose, it was necessary to intervene on the treebank by marking complements, sentences with null or inverse subject, and so on.

The preprocessing task accomplished by Corazza et al. is summarized here below and is actually restricted to the use of lemmas in place of word forms as head of lexicalized constituents (see ibid., pag.4). From the verifications carried out using two different parsers, researchers have come to the conclusion that,

> "These preliminary results... confirm that performance on Italian is substantially lower than on English. This result seems to suggest that the differences in performance between the English and Italian treebanks are independent of the adopted parser... our hypothesis is that the gap in performance between the two languages can be due to two different causes: intrinsic differences between the two languages or differences between the annotation policies adopted in the two treebanks."(p.5-6)

From the experiment computed on the basis of the information theory it turns out that the difference in performance cannot be imputed to the amount of rules and therefore to the type of annotation introduced, but to the scarce predictability of their structural relations, as stated by the authors,

> "First of all, it is interesting to note how the same coverage on rules results in the Italian corpus in a sensibly lower coverage on sentences (26.62% vs. 36.28%). This discrepancy suggests that missing rules are less concentrated in

the same sentences, and that, in general, they tend to be less correlated the one with the other. This would not be contradicted by a lower entropy, as the entropy does not make any hypothesis on the correlation between rules, but only on the likelihood of the correct derivation. This could be a first aspect making the ISST task more difficult than the WSJ one. In fact, the choice of the rules to introduce at each step is easier if they are highly correlated with the ones already introduced."(p. 9)

2.1 Regularity and discontinuity in the language and its linguistic representation

A number of conclusions can safely be drawn from what the researchers stated and from the results of their test. Intuitively one could assert that the better the structural regularity of a language or its representation is, the wider its reproducibility on a statistical basis; on the contrary, in a language containing many cases recurring only once, in general hapax, bis-, tris- legomena, a good statistical result of the model is less probable – this is what is usually referred to as the sparcity/sparseness problem.

In linguistic terms the issue can be due to the division of grammar into core and periphery and this partition should be characterized in a quantitative manner. A statistical parser needs a great number of canonical structures belonging to the core grammar and it is not a case that in his procedure of creation of the model Collins deliberately introduces some corrections in the original treebank; that is, one has to accurately account for the structures which compose the core grammar, while the ones that constitute the periphery are amended ad hoc. Therefore, the malfunctioning of a statistical parser trained on a treebank must be related to the reference linguistic framework chosen by the annotators and hence to the reference language.

From the global quantitative data reported in **Table 2.** below, one can see that much more than half of the Italian (simple) sentences (9.800 in 19.099) do NOT have a subject lexically expressed in canonical position: this makes it very unpredictable to locate the SN Subject. If we compare this with PT we get a completely different picture. Regarding non-canonicity, in PT there are 4647 sentences which have been classified with the node of topicalized structure (S-TPC) a label which includes argument preposing, sentences in direct reported speech, and so on. Moreover there are sentences with an inverse structure, classified as SINV, only 827 of which are also TPC. SINV sentences are 2587 and they all typically have the subject in post-verbal position.

While as for the work on PT it is sensible to correct the problem in the pre-processing phases as made by Collins and commented by Bikel, in our case this issue is much less sensible and certainly more complicated to carry out. In fact, the Subject NP can be realized in four different ways: it can be lexically omitted – this constitutes the majority of cases as discussed below-; it can be found in an inverted position in the COMP(T/IN/C/PASS) constituents where complements are placed; it can be found in dislocated position on the left or on the right of the sentence to which it is related, at CP level. In a preliminary annotation of such cases we counted a total of more than 3000 cases of lexically expressed subject

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

in non-canonical position. Then there are about 6000 cases of omitted subject to be taken into account. All these sentences must be dealt with in different ways during the creation of the model as discussed below.

If one considers that in PT there are 93532 sentence structures – identifiable with the reg_ex "(S (" - 38600 of which are complex sentences, that is the 41% of all the "(S (" – adding up all the cases of non-canonical SUBJect sums up to a very low percentage, around 1%. On the contrary, in VIT the same phenomenon has a much higher import, over 27% in the case of non-canonical structures, and over 50% as to the omitted or unexpressed subject. We have also taken into consideration the annotation of complements in non-canonical position, and they have been listed in a table below.

Treebanks Vs. Non-canonical Structures	**Non-canonical Structures (TU)**	**Structures with Non-Canonical Subject (TS)**	**Total (TU) Utterances**	**Total (TS) Simple Sentences**	**Totale Complex Sentences**
VIT	**3719**	**9800**	**10,200**	**19,099**	**6782**
Percentage	**27.43%**	**51.31%**	**63.75%**		**66.5%**
PT	**7234**	**2587**	**55,600**	**93,532**	**38,600**
Percentage	**13.01%**	**0.27%**	**59.44%**		**69.4%**

Table 2. Comparison of non-canonical Structures in VIT and in PTB where we differentiate TU (total utterances) and TS(total simple sentences)

Here below in Table 3. we show absolute values for all non-canonical structure we relabeled in VIT. Considering that the total number of canonical lexically expressed SUBJects is 7172, we can compute the number of non-canonical subjects as constituting 1/3 of all expressed SUBJects – total number of lexically expressed subjects corresponding to 10,100. We labeled as S_TOP subject NPs positioned to the right of the governing verb; as S_DIS those subject NP which are positioned to the left of the governing verb but are separated from it by a parenthetical or a heavy complement; S_FOC are typically subject in inverted postverbal position of presentational structures; finally LDC are all types of Left Dislocated Complements with or without a doubling clitic.

	LDC - left dislocated complements	S_DIS dislocated subject	S_TOP topicalized subject	S_FOC Focalized Subject	Total Non-Canonical	Totale Complex Sentences
VIT	**251**	**1037**	**2165**	**266**	**3719**	**3039**

Table 3. Non-canonical Structures in VIT

2.2 Discontinuous modification

More problematic syntactic structures at all those structures that in Italian can have a modification or argument role in nominal structures, some of which can be found either before or after the head and some others can be dislocated in a distant position – separated by other intermediate constituents – in particular we here are referring to adjectival structures that can freely occur in post-nominal position even at a remarkable distance from the head – this structure is not possible in English. The relative data are reported in **Table 4.** where we counted the frequency in terms of the distance represented as the number of square brackets from the closest head, always a nominal head. In the case of a complement (argument) the constituent would be adjacent to the head otherwise it would be separated by a certain number of brackets that varies from 1 to 4.

A first reading gives us quite intuitive data as regard to the role that these constituents have in sentence structures: in particular the Ratio AM/AC tells us how many constituents there are of a certain type that have the function of argument/adjunct at sentence level compared to those that have the function of modifier in nominal structures. As one can see, PP-OF and RC are the two structures that more than others can be found in nominal structures, on the contrary the PP-BY rarely takes this role. Moreover, PP-OFs and APs differentiate themselves roughly from all the other constituents as to the argument position we named "Head Adjacent". It is important to note how, in the case of APs, 8169 constituents (35%) are actually in a pre-head position although calculated as "Head Adjacent". PP-OFs and RCs in post-head nominal position are respectively about 90% and 73% of all this type of constituent. As for RCs, 845 of them (25%) are of the non-restrictive type, i.e. they are separated by a comma: 98 of these are separated from the nominal head by a modifier, while the remaining ones present a PP type structure or a deeper embedded structure between the head and the RC whose dependency is thus difficult to identify. In the case of PPs and PP-BYs they are respectively 48% and 38% of the total amount of occurrences: in fact, in most cases these constituents have the function of complement/argument and of adjunct located in the COMP structure or at CP level. Lastly, 65% of the VPs are distributed as modifiers, while in the remaining cases they can occur as SUBJect of copulative sentences or as argument in the COMP structure.

It is easy to guess that the constituents with a higher structural ambiguity in Italian are those whose position in respect to the head is less predictable: respectively AP>VP>PP>RC>PP-BY>PP-OF. Besides, we must consider other elements that can lead to discontinuity or non-canonicity problems. In particular, the number of F3 or sentence fragments is quite high compared to the number of total utterances, 3552 (35%); the number of complex utterances is quite high – 6782 if compared to the total number (10.200) of utterances, therefore much higher than the 41% of PT.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Constituent/ Distance	SP	SPD	SPDA	SV	F2	SA	TOTAL
Head Adjacent (HA)	4726	13.798	509	3249	1560	13.932	37,774
Distance=1	2677	1827	266	941	460	908	7,079
Distance=2	1718	494	203	485	305	179	3,384
Distance=3	624	81	58	130	82	24	999
Distance=4	600	45	32	175	100	23	975
Total All Mods (AM)	10.345	16.245	1068	4980	2507	15.066	50,211
Ratio AM/AC	***0,483***	***0.912***	***0.384***	***0.658***	***0.73***	***0.71***	***0.652***
Totals Non HA	5619	2447	559	1731	947	1134	12,437
Ratio Non HA/AM	***0,54***	***0,15***	***0,523***	***0,347***	***0,378***	***0,075***	***0.652***
All Constituents	21.393	17.812	2780	7568	3425	21.205	76,971

Table 4. Comparative Data of the position, in relation to the head, of the constituents that can be modifiers in nominal structures. See Table 1 for the list of the constituents. AM=all modifiers; AC=all constituents; HA=head adjacent

3 Ambiguity and Discontinuity in VIT

We will briefly present and discuss some of the most interesting structures contained in VIT as regards the two important question of ambiguity and discontinuity in Italian. The most ambiguous structures are constituted by Adjectival related structures. As already commented above, adjectives in Italian may be positioned in front or after the noun the modify almost freely for most lexical classes. Only few classes require to be in predicative position and a very small number of adjectives must be placed in front of the noun they modify, in attributive position. A count of the functional conversion of adjectival structures is presented here below:

1296 Complement APs (ACOMP), 18748 Modifiers (MOD), 324 Adjuncts (ADJ), 2001 COORDinate APs

Postnominal adjectives constitute the most challenging type since they may be considered as either post or premodifiers of a following nominal head. Even though postnominal non-adjacent SA recur in a small number – only 5.34%, they need to be identified by the parser. In the examples below we try to show how this process requires knowledge of adjectival lexical class besides feature matching. For every example taken from VIT we report the relevant portion of structure and a literal translation.

(1) sn-[art-i, n-posti, spd-[partd-della, sn-[n-dotazione, sa-[ag-organica_aggiuntiva]]], sa-[ag-disponibili, sp-[p-a,
the posts of the pool organic additive available to

Syntactic ambiguity arises and agreement checking is not enough even though in some cases it may solve the attachment preferences for the predicative vs. the attributive position.

(2) sn-[sa-[ag-significativi], n-ritardi]], sn-[sa-[ag-profonde], n-trasformazioni], ibar-[vt-investono],
significative delays profound transformations affect

Adjectival structures may come in a row and modify different heads as in,

(3) sn-[art-il, n-totale, spd-[partd-dei, sn-[n-posti, spd-[partd-della, sn-[n-dotazione, sa-[ag-organica]]], ag-vacanti], sa-[ag-disponibili
the total of the posts of the pool organic additive vacant available

where "vacant" modifies the local head "posti", as well as "disponibili" which however governs some complement. On the contrary, in the example below, "maggiori" is not attached to the a possible previous head "orientamenti", but to a following one as the structure indicates,

(4) ibar-[vin-darebbe], compin-[sp-[in-anche, part-agli, sn-[n-orientamenti, spd-[pd-di, sn-[n-democrazia, sa-[ag-laica]]]]], sn-[sa-[ag-maggiori
would give also to the viewpoints of democracy laic main

Another interesting phenomenon present in Italian is the possibility to have fronted PP complements in Participials. This structure may cause ambiguity and problems of attachment, as shown in the examples below,

(5) sp-[p-in, sn-[n-base, sp-[part-al, sn-[n-punteggio, sv3-[sp-[p-ad, sn-[pron-essi]], ppas-attribuito, compin-[sp-[p-con,
on the basis of the scoring to them attributed with

where we see that "ad essi" could be regarded as a modifier of the previous noun "punteggio", whereas it is a complement of "attribuito" which however follows rather precede it.
Another more complex case is constituted by,

(6) sp-[p-a, coord-[sn-[sa-[ag-singoli], n-plessi], cong-o, sn-[n-distretti], sv3-[sp-[p-in, sn-[pron-essi]], ppas-compresi, punto-.]]]]]]]]]
to single groups or districts in them comprised

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

As the examples show, this could also be computed as a case of proclitic, seen that only personal pronouns are allowed to be fronted and not nouns,

(7) spd-[partd-degli, sn-[n-importi, sv3-[sp-[p-ad, sn-[pron-essi]], ppre-spettanti]]], cong-e,
of the amounts to them owed and

The structure is not only found in bureaucratic language but also in literary genre, as in,

(8) spd-[partd-della, sn-[n-cortesia, sv3-[sp-[p-in, sq-[q-più, pd-di, sn-[art-un_, n-occasione]]], vppt-dimostrata, compin-[coord-[sp-[p-a, sn-[pron-me]],
of the courtesy in more than one occasion demonstrated to me

Other non canonical structures are constituted by Subject Inversion, Focus Inverted APs, Left Clitic Dislocation with Resumptive pronoun that we don't have the space to show here.

Finally we will present and discuss some Aux-to-comp structures attested again both in bureaucratic and literary genres.

(9) topf-[auxtoc-[auag-avendo], f-[sn-[art-la, npro-Holding], sv3-[vppt-incassato, compt-[sn-[n-indennizzi, sp-[p-per, sn-[num-'28', num-miliardi]]]]]], punto-.]
having the Holding cashed payments for 28 billions

Here the gerundive auxiliary precedes the subjec NP which in turn precedes the lexical verbal head in participial form. Below is a typical only Italian aux-to-comp structure,

(10) fc-[congf-e, punt-',', topf-[auxtoc-[clit-si, aueir-fosse], f-[sn-[pron-egli], sv3-[vppin-trasferito, cong-pure, compin-[sp-[part-nel, sn-[sa-[in-più, ag-remoto], n-continente]]]]]
and , self would be he moved also in the more remote continent

This case and the following only belong to literary genre,

(12) cp-[sn-[topf-[auxtoc-[art-l, ausai-avere], f-[sn-[art-il, n-figlio], sv3-[vppt-abbandonato, compt-[sn-[art-il, n-mare], sp-[p-per, sn-[art-la, n-città]]]]]], f-[ibar-[clitdat-le, ause-era, avv-sempre, vppt-sembrato]
the have the son abandoned the sea for the city her was always seemed

Peculiarities in common with classical aux-to-comp is the presence of an auxiliary as structural indicator of the beginning of the construction. We introduced a new special constituent TOPF to include the auxiliary and the sentence where the lexical verbal head has to be searched in order to produce an adequate semantic interpretation.

4 Conclusions

VIT differs greatly from PT not only for the amount of sentences and data, but also for the choice to include linguistic material of different nature: in VIT there are five different genres – news, bureacratic genre, political genre, scientific genre, literary genre -, while in PT only one is represented. Hence the wider homogeneity we expect from PT and consequently the scarcer homogeneity in VIT.

The sparcity of VIT makes it difficult, if not impossible, to use it as a Language Model in the construction of probabilistic grammars for Italian. Therefore it is necessary to introduce corrective elements in order to enable the learning phase to distinguish sentences with different typologies (subject in canonical preverbal position, subject in non-canonical post-verbal position, lexically unexpressed subject, left dislocated "hanging Topic" subject – separated from the verb by other complements (or composed of a "heavy" SN followed by punctuation) - right dislocated Hanging Topic subject – separated from the verb by other complements), etc. . To this end, we are implementing Bikel's language model directly on VIT and from preliminary results we can safely say that the same poor performance is reconfirmed – around 70% accuracy. More experiments will be carried out to confirm the hypothesis in Corazza et al., even though from the data in our possession such a confirmation is very likely.

References

Bikel, Daniel M. 2003. Intricacies of Collins' parsing model. Computational Linguistics, **30**(4), pp. 479-511.

Corazza A., Lavelli A., Satta G., Zanoli R. 2004. Analyzing an Italian Treebank with State-of-the-Art Statistical Parsers**.** In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT-2004*), pp. 39-50, Tübingen, Germany.

Delmonte, R. 2004. Strutture Sintattiche dall'Analisi Computazionale di Corpora di Italiano, in Anna Cardinaletti(a cura di), *Intorno all'Italiano Contemporaneo*, Franco Angeli, Milano, pp.187-220.

Delmonte R. 2000. Shallow Parsing And Functional Structure In Italian Corpora, LREC, Atene, pp.113-119.

Gildea D. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pp. 167–202, Pittsburgh, PA.

Montemagni, S. F. Barsotti, M. Battista, N. Calzolari, A. Lenci O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Basili R. Raffaelli, M.T. Pazienza, D. Saracino, F. Zanzotto, F. Pianesi N. Mana, and R. Delmonte 2003. Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé, editor, *Building and Using syntactically annotated corpora*, pp. 189–210. Kluwer, Dordrecht.

Musillo G. & Khalil Sima'an 2002. Towards comparing parsers from different linguistic frameworks. An information theoretic approach. In *Proceedings of the LREC-2002 workshop Beyond PARSEVAL. Towards Improved Evaluation Measures for Parsing Systems*, pp. 44-51, Las Palmas, Spain.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476