Syntactic and Lexical Complexity in Italian Noncanonical Structures

Rodolfo Delmonte
Università Ca' Foscari
Ca' Bembo, Dorsoduro 1745, 30123 - VENEZIA
E-mail: delmont@unive.it - website: project.cgm.unive.it

Abstract

In this paper we will be dealing with different levels of complexity in the processing of Italian, a Romance language inheriting many properties from Latin which make it an almost free word order language[1]. The paper is concerned with syntactic complexity as measurable on the basis of the cognitive parser that incrementally builds up a syntactic representation to be used by the semantic component. The theory behind will be LFG and parsing preferences will be used to justify one choice both from a principled and a processing point of view. LFG is a transformationless theory in which there is no deep structure separate from surface syntactic structure. This is partially in accordance with constructional theories in which noncanonical structures containing non-argument functions FOCUS/TOPIC are treated as multifunctional constituents. Complexity is computed on a processing basis following suggestions made by Blache and demonstrated by Kluender and Chesi.

1 Introduction

In this paper we will be addressing what the CFP of the workshop has defined as "whether, and to what extent, linguistic phenomena hampering human processing correlate with difficulties in the automatic processing of language". This will be done by presenting work done in the past on the topic of noncanonical and difficult to parser syntactic structures that may create ambiguity at phonological level, and how to solve it. In that case, the goal was creating an automatic system for text-to-speech, i.e. a TTS system. This will be completed by showing results of an experiment done with statistical data-driven dependency parsers - and a rule-based one - analysing a highly noncanonical text type – children stories or fables. I will show the evaluation done both on the Italian text and the corresponding English translation.

Current approaches to deep syntactic-semantic natural language modeling are strongly statistically based and have achieved near 90% accuracy in some cases. The problem with statistical modeling is that they are strictly bound to training material. Achieving generality requires mixing diverse domains in the training data. In such cases, accuracy varies a lot depending on the language. In particular, more canonical languages achieve 85% accuracy on average – this includes English in non-projective structures and 90% accuracy on projective ones[2]. Less canonical languages, like Italian for instance, are below that threshold and average 82/83% accuracy[3].

The question is that syntactic processing is just one step towards natural language understanding, and the hope to cover sentence level semantics in the near future is not very close. In addition, if we consider Italian, the results are strongly flawed by the fact that dependency parsers are not equipped for lexically unexpressed categories as the Null Subject, which in Italian constitute some 60% of all Subject cases. Languages containing Null Subjects also include Chinese where the recognition rate for this language of such Subject positions by current state-of-art statistical parsers averages 50%

[1] Parameters usually referred to when defining "free" word order languages like Latin include: lack of articles, Null Subjects, lack of expletives, freely omitting the complementizer, intensive case marking, etc. to quote the most important ones. Italian only has some of them and the resulting constructions for a simple declarative clause may include all possible permutations at constituent level, but not at word level as Latin for instance would do.

[2] Experiments with different domains for test and training are reported in Surdenau et al. 2008 where in particular the conplete task with WSJ(Marcus et al. 1993) averages 86% F measure, when Brown(Francis & Kucera 1967) is used it drops to 76%. Average performance for 5 different domains when training and test domain diverge show a significant drop whenever we move from WSJ 89% to biomedical domain GENIA 66.6%, to dialogue domain of SwitchBoard 69% and to Brown 80%.

[3] This value has been reported in a mail thread by Giuseppe Attardi experimenting with Universal Dependencies treebanks and the newly released SyntaNet, "the world's most accurate parser", as it is publicized on the web.

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 67–78, Osaka, Japan, December 11-17 2016.

accuracy. Semantic processing is thus highly flawed by the grammatically incomplete structures produced by data-driven dependency parsers.

An important factor that determines levels of complexity in linguistic data are presence of non-canonical structures which in some languages and some domains are almost negligible, as for instance English in the corpus constituted by WSJ news – see below. But when we move to the BROWN corpus the presence of such structures is important and determines a decrease in accuracy and a drop in performance that can range from 6 to 8 points (See McClosky et al. 2010, Hara et al. 2010, Gildea 2001). Italian on the contrary is very rich on such non-canonical structures including discontinuities of all sorts, but as before some genre or domain has more than others.

The paper is organized into two main sections, the following one, section 2 devoted to performance related cases of complexity where we take the stance to use LFG theory and Parsing Strategies to explain ambiguity in the data. We will then present results of a study carried out on non-canonical Italian sentences as they have been treated by most well-known parsers of Italian. In this section we will then show results from an experiment with current best statistical data-driven dependency parsers of Italian when presented with a highly non-canonical text.

2 Different types of complexities: Performance and Parsing Strategies

As the call for paper clarifies, the notion of complexity is highly polysemous and may be related to a number of different issues. In the past, we have been approaching language complexity by way of structural ambiguity and parsing preferences, in a number of papers dealing with some of the issues mentioned above that we will present shortly below. In particular, in the paper published in Delmonte, 1984 we discuss issues related to so-called "Syntactic Closure" where performance and theoretical issues are strongly interrelated. Similar problems have been discussed in the paper published in Delmonte, 1985 which poses questions linking phonology and parsing, again performance related issued.

In our approach we have always indicated LFG (Lexical-Functional Grammar)(Bresnan, 1982) as the theoretical and practical backbone of our research activity, which has inspired also heavily the way in which our computational work has been carried out. This is due to the choice of LFG to support a psycholinguistic approach in which performance played an important role and to call for a processor as a fundamental component of the overall theory(ibid. xxii-xxiv). LFG is a transformationless theory which is based on the lexicon and the existence of lexical rules to account for main NP movements. Long-distance dependencies are accounted for by properties of the f-structure. In a paper appearing in 1989, Kaplan & Zaenen introduce the notion of "uncertainty" in functional assignment. A displaced f-structure, receiving one of the non-argumental pragmatically related functions, FOCUS or TOPIC, will be made dependent or will fuse with the missing function in a following or preceding f-structure that requires it. Requirements are dictated by grammatical principles of Completeness and Uniqueness. Lately, long-distance dependencies in LFG and processing related theories have been pragmatically based in a paper by Y.Falk(2009) and we will return to this position in the final section.

2.1 Syntactic Closure and Parsing Strategies

The problem of Syntactic Closure (hence SC) will be here discussed inside a Theory of Performance or a Linguistic Realization Theory(hence LRT) which in turn is represented in LFG. Consider the two examples (ibid. 103, same numbering):

(1) Tino ha detto che Bruno è morto ieri. / Tino said that Bruno died yesterday.
(2) Tino ha detto che Bruno è partito ieri. / Tino said that Bruno left yesterday.

A competence theory will give both sentences an ambiguous structural description, on the basis of the fact that the adverbial "ieri/yesterday" can be attached both in the embedded and in the main clause, and will make available two structural representations. The theory of SC (hence TSC) which operates inside LRT has to explain why there is a preference for a particular analysis, i.e. attaching the adjunct lower, inside the embedded clause and not in the clause governed by the main verb DIRE/SAY. In LFG Adjuncts are not argument-like structures and cannot be subcategorized in the lexicon. There can only be semantic restrictions for compatibility and appropriateness.

Syntactic ambiguity can disappear in case TENSE in the embedded clause is no longer "compatible" with tense features contained in the Temporal Adjunct, as for instance in the following examples (ibid. p.104, same numbering):

(3) Giovanni ha detto che non tornerà più ieri. / John said he will never come back yesterday
(4) Ho trovato le scarpe che più desidero ieri. / I have found the shoes I like more yesterday

However, there may be still another class of examples, this time lexically biased, where tense is compatible but attachment will preferentially go up to the main clause (ibid. p.105, there (8) and (9)):

(5) Finalmente ho terminato il lavoro che mi rendeva schiavo ieri. / At the end I finished the work which made me slave yesterday.
(6) C'è che ho lasciato la persona che amavo ieri. / It happened that I left the person I loved yesterday.
(7) Sai, ho trovato il regalo che speravo ieri. / You know what, I found the present I hoped yesterday.

In these examples, the strong form of the verb attracts attachment, the weak form doesn't. The three verbs are lexically represented as follows according to LFG:

TERMINARE	<(SUBJ), (OBJ), (ADJ)>
AMARE	<(SUBJ), (OBJ)>
SPERARE	<(SUBJ), (XCOMP)>

The attachment attraction is induced by the force of alternative lexical entries for the same verb, and this in turn is determined by frequency of usage and a stability in the underlying structure that justify these findings on SC also in sentences isolated from context (p.105). As said above, semantic appropriateness depends on the knowledge of the world the speaker will have available in order to decide which sequence of predicate, argument sand adjuncts to select in the context.

The theory of sentence understanding that we purport assumes that the order of application of the rules in a rule-based parser is determined by two principles obeying certain parameters. The two principles are Lexical Preference (LP) and Final Arguments (FAs); the former has as default parameter the strength of alternate categories in the choice of syntactic rules called Syntactic Preference(Delmonte, 2000). The latter principle, FAs, is bound to the hypotheses developed by the parsing process, and the parameter is called Invoked Attachment (IA). The "closure" property of syntactic phrases in turn is governed by lexical elements and their underlying lexical form, as indicated by Ford et al., (1981, 747). The default principle to prioritize alternate categories in the execution of phrase structure rules is the strength of those categories (ibid. 749). In the two sentences (our numbering, p.111, there 18, 19):

(8) La 1 donna 2 ha 3 sistemato 4 il 5 vestito 6 su 7 quell' 8 attaccapanni 9 . 10 / The woman has hanged the dress on that peg.
(9) La 1 donna 2 ha 3 chiesto 4 il 5 vestito 6 su 7 quell' 8 attaccapanni 9 . 10
/ The woman has asked the dress on that peg.

As the parser reaches position (4) it has to build the OBJect and then complete the VP. To do that, it will call for lexical information, which have the following strong forms:

SISTEMARE	<(SUBJ), (OBJ), (PCOMP)>
CHIEDERE	<(SUBJ), (OBJ)>

After choosing to build an OBJect, at position 6, in one case, - with the verb SISTEMARE/HANG - the LP principle will hamper taking the PP to be analysed as ADJunct of the noun DRESS and so consumed locally inside the OBJect NP. The parser will require the PP to be interpreted as PCOMP and as such be left for the upper VP level. This will also be promoted by the other principle, FAs, which says that: the final argument in the lexical form of a given predicate is a syntactic phrase that must be coherent with it - NP/OBJ, PP/PCOMP - and cannot be followed by other constituents

coherent with the same form. On the contrary, the LP will allow the PP to be consumed locally in the OBJect NP. In one case the theory speaks of LATE CLOSURE, where the PP is prevented from being consumed inside the NP headed by DRESS. The PP in this case has been given low priority by the presence of alternate options, i.e. the presence of a PCOMP in the strong form of the main governing verb SISTEMARE/HANG.

2.2 Functional vs Syntactic Reversibility in Question Parsing and Semantic Roles

In Delmonte(1985) we were basically concerned with a recognition grammar to supply information to a text-to-speech system for the synthesis of Italian which is shown to have to rely heavily upon lexical information, in order to instantiate appropriate grammatical relations and assign Semantic Roles. Italian is an almost free word order language which nonetheless adopts fairly analysable strategies for major constituents which strongly affect the functioning of the phonological component. Two basic claims have been made: i. difficulties in associating grammatical functions to constituent structure can be overcome only if Lexical Theory is adopted as a general theoretical framework, and translated into adequate computational formalisms like ATN or CHART; ii. decisions made at previous point affect Focus structure construal rules, which are higher level phonologicaI rules that individuate intonation centre, contribute to build up adequate Intonational Groups and assign pauses to adequate sites, all being very sensitive to syntactic and semantic information. In the paper, I then concentrate on Subject/Object function association to c-structure in Italian, and its relation to ATN formalism, in particular HOLD mechanism and FLAGging. This is done by analysing wh- structures, both direct questions and relative clauses which eventually constitute non-canonical structures in which Subject Inversion is almost obligatory.

We define reversible structures at syntactic and functional level as the ones that allow their arguments to assume both SUBJect and OBJect functions as in the examples below (ibid. 137, number 1,2):

(12)a The secretary has been killed by the director.
b. The book has been read by John.

While 12a. allows reversing the two core functions, the b. example doesn't. It is clear that non-reversible passive structures contain additional grammatical cues to speed up comprehension, which are only available from lexical entries in which selectional restrictions are listed. These features are then used to constrain assignment of semantic roles. From a purely processing point of view, passive structures are the canonical case of NP requiring reinterpretation when verb morphology is accessed. Thus, the NP SUBJect computed so far, will receive reversed Semantic Role associated to the OBJect in the lexical entry of the verb – in LFG by means of lexical rules.

As said above, Italian is a language that allows Null Subjects: SUBJects in Italian have specific properties:

- it can appear in preverbal – the canonical position - or postverbal position as a case of Subject Inversion;
- be unexpressed as a case of obviative or extrasentential pronominal in tensed clauses;
- be stranded or extraposed, i.e. moved out of its matrix clause and placed after heavy Complements (phrases or sentences)

It is also necessary to clarify that lexical properties are paramount also in English, where not always NP1 appearing in preverbal position entertains SUBJect function, nor NP2 can be always interpreted as OBJect as the following examples clearly show (ibid. 137, numbers 3-7)[4]:

(13)a. Computers have been given no consideration whatsoever by linguists in Italy
b. Her father Mary hates.

[4] where we have cases of fronted NP2 detectable only by having access to NPs inherent semantic features. Thus, in a.,it is OBJect2 which has been passivized and not NP2; in b. we have a topicalized sentence with fronted NP2; in c. SELL is used in ergative structural configuration, in which NP2 is raised to Subject; the same applies to d., a case in which Subject NP would be always omitted (subjectless impersonal structures are frequently used in technical and scientific English); also e. is a subjectless structure, in which "tough predicate" appears and Object NP2 is raised to Subject position.

c. The latest book by Calvino sells well.
d. The logical operator .NOT. applies to the parenthesized statement.
e. Geneva is easy to reach in Italy.

And now briefly, NP2 need not always be interpreted as Object of its clause, as shown below (ibid. 137, numbers 8'10)[5]:

(14)a. There came the magician with his magic rod.
b. But the real murderer is the landlord.
c. Mary gave John a beautiful present.
d. In the corner stood an old boxer.

Now, STRUCTURAL reversibility involves the possibility to use the same constituent order and to freely alternate the instantiation of grammatical functions, while the underlying Semantic Roles change. The result is that with Structural reversibility only one interpretation will result. Even though Semantic Roles can be associated interchangeably to either preverbal or postverbal NP without violating selectional restrictions or semantic compatibility conditions, it is the final constituent order and structure that decides on the interpretation. In this sense, non-reversible passives only allow a single well-formed mapping.

Coming to wh- structures we can see the difference existing between Italian and English: the following example is only allowed in Italian where in fact it is obligatory (ibid.138, number 13):

(15)a. *This is the cheese that ate the mouse that ate the cat that chased the dog.
b. Questo è il formaggio che ha mangiato il topo che ha mangiato il gatto che inseguì il cane.
c. Questo è il formaggio che ha mangiato il topo che ha mangiato il gatto che il cane inseguì.
d. This is the cheese that the mouse ate that the cat ate that the dog chased.

This example shows a case of non-reversible functional structure: postverbal positions are available structurally but not functionally in English which semantically will only allow SUBJect interpretation for MOUSE, CAT, DOG but require preverbal positioning. Italian makes available postverbal position by means of SUBJect Inversion and thus the correct interpretation is triggered by the lexicon. These sentence contain double non-canonical structures – relatives and inversion - and require more computation than canonical ones. The canonical version would have the relative pronoun for CHEESE interpreted in OBJect position, then a preverbal SUBJect position for the DOG.

In STRUCTURAL reversibility constituent order is crucial and characterizes configurational languages with fixed word order. On the contrary, with FUNCTIONAL reversibility ,constituent order is irrelevant, and what really matters is a lexically informed and constrained mapping. In configurational languages, grammatical functions can be associated in a reliable way to fixed or canonical constituent orders - examples 13. 14. above are both structurally and lexically marked. In Italian no such order exists because both preverbal and postverbal constituent positions constitute an unmarked case for SUBJect/OBJect functional assignment.

As a result, a parser of Italian is unable to produce reasonable predictions on the underlying grammatical relations in lack of morphological and lexical cues: it will have to rely on lexical and extralinguistic information. To better exemplify this, we will discuss wh- constructions which in English are more easily computable but in Italian are usually difficult to parse. The following example is very instructive (taken from Ritchie, 1980, his 120):

(16)a. Dove ha sepolto il tesoro che ha rubato l'uomo di cui parlavi?
b. Where did the man who you mentioned bury the treasure which he stole?
c. *Dov'è che l'uomo che hai menzionato ha sepolto il tesoro che ha rubato?

[5]where a. is a presentation sentence with a dummy pronoun "there" and the Subject NP is in postverbal position; b. is a predication sentence in which something is predicated about the NP Subject "the landlord" in postverbal position; in c. the postverbal NP is OBject2 of ditransitive Verbs constructions, which has undergone dative shift; and in d. we have a case of locative inversion.

In a. the NP SUBJect "l'uomo" has been displaced beyond two bounding nodes - in Italian NP and S' count as such (Rizzi, 1980): it binds two SUBJect positions and also the NP OBJect position inside the lower relative clause headed by "cui". On the contrary, the Null Subject position in front of "parlavi" is assigned obviative or disjoint reference, to an external antecedent. The only correct version for the English example reported in b. is a. On the contrary c. which tries to translate literally the same order is ungrammatical. The same happens with yes-no questions like this one (ibid.139, number 15):

(17)a. Ha finito i compiti tua sorella?
b. Has your sister finished her homework?

where postverbal position is again reserved for NP Object and the NP Subject "tua sorella" has been stranded or "extraposed". Simple wh- questions have the same problem, i.e. they lack structural cues to help detecting functional assignment, as in examples below (ibid. 139, numbers 16,17):

(18)a. Quale pesce ha pescato la segretaria?
b. Quale segretaria ha pescato il pesce?
c. Which fish did the secretary catch?
d. Which secretary caught the fish?

Fully ambiguous structures are the following complex Italian wh- questions(ibid.139, numbers 19,20):

(19)a. Chi era la persona che ha incontrato Gino?
b. Who was the person who met John?
c. Who was the person who John met?
(20)a. Chi ha detto che avrebbe assunto il capo?
b. Who said that he/she would have hired the chief?
c. Who said that the chief would have hired?

Both INCONTRARE and ASSUMERE are only transitive verbs that make available two NP positions which are fully functionally reversible. On the contrary PESCARE that we saw before, is not functionally reversible. Another possibility would be the one constituted by the example below (ibid. 140, number 24):

(21)a. Chi ha detto che sta arrivando Gino?
b. Who said that Gino is arriving?

In this case, the only functional role that can be associated to "Gino" is the one of inverted SUBJect seen that ARRIVARE is an unaccusative verb. Eventually we present a case of passive focussing, which is in many cases obligatory – as this one - in order to convey the novelty of the event:

(22)a. E' stato ucciso il presidente Kennedy! /President Kennedy has been killed.
b. Il presidente Kennedy è stato ucciso./ President Kennedy has been killed.

Example b. is a version of an agentless passive in which the SUBJect is already Topic of discourse and the news needs only be confirmed. On the contrary, 22a presents the news out of the blue. The question here is complicated by the fact that the SUBJect is postponed in OBJect position to convey new information. In a transformation grammar, this would require movement back and forth, twice: at first to recover the grammatical function of sentential SUBJect, and then back since the same NP constituent has to be interpreted as Affected Theme and not as Agent in deep structure. Example 22a. is impossible in English, and only 22b is allowed: linguistic theories have been using Passive structures to determine their difference and specialty in the treatment of this important structure. They have all been based on English examples: Subject inverted structures in passive constructions might have induced different theoretical approaches. Similar to Italian examples are Russian Subject inverted structures, which have been considered as strictly depending on information structure distribution, due

to contextual factors: Partee et al. 2011 treats these cases as Perspective Structures. Subject inversion in Russian involves Existential sentences, Locative inversion, Passive inversion but also Unergative inversion (see Glushan & Calabrese, 2014).

3 Computing Complexity

In the ATN formalism, examples 19/20 are analysed as follows: a question element is contained in a register HOLD which is used to store it temporarily until the rest of the clause is processed. Then the element is passed down to any constituent that might use it - NP SUBJect/OBJect - or in turn could be allowed to pass it down to one of its internal constituent in case of complement clauses. Eventually, CHI might be made to fill in the lower SUBJect position or even the lower OBJect position, but in this case, "il capo" should be made to climb up – or erase and substitute the contents of the register by taking the position now occupied by "little_pro", a certainly more expensive choice. In a CHART inspired parse of the same sentences, all structures would be made available and the decision to choose the most appropriate would be left for the discourse level to make. The HOLD mechanism does not seem to be particularly adequate to solve the ambiguous structures we proposed, since it usually works searching for a HOLE where to insert some linguistic material it has already found and judged to be displaced. In our case the situation is totally reversed: first come the HOLE(S) then the material to fill in. To suit the limitations imposed by Short-Term or working Memory, no more than 7 single linguistic items can be stored before they enter Long-Term Memory. This is mimicked by the working of an incremental parser that computes fully interpreted structures which in LFG should correspond to F-structures. In our previous work(2009, Chapter 3) we presented a principle-based version of the parser that takes advantage of the Minimalist Theory (hence MT) to instruct the "processor" while inputting words incrementally in the working memory. In the sections above, the theory we followed was LFG but in both cases it is now the lexicon that drives the computation: in fact, lexical information in the MT makes available features to the processor that will use Merge and Move to select appropriate items and build a complete structural representation. To justify memory restrictions, the parser takes advantage of an intermediate level of computation, called PHASE constituted by a fully realized argument structure preceded by a Verb. Reaching the verb is paramount also for LFG theory, for selection but also for grammatical function assignment. In Chesi(2016:31-32) the author proposes a complexity metrics called Feature Retrieval Cost (hence FRC) that he associates to the MT theory. In particular, the moves of an MT-inspired processor are accompanied by reading times, which seem to (partially) confirm the prediction of the theory. In turn, these predictions are computed on the basis of local and non-local features and depend strictly on the type of argument selection operated by the verb. This is what has been discussed above, with the so-called REVERSIBILITY notion applied to structure and function.

In addition, Chesi(ibid.33) following Friedmann et al. and Gibson(1998), assumes that referential properties of NPs as shown at determiner level may induce different cost measures: definite NPs being heavier to process than Proper Nouns, and these in turn heavier to process than deictic personal pronouns (you/I). This hierarchy is then partially reinforced by presence of distinct features. Reversibility however is a cost inducing factor, but as we saw above, it applies whenever strong transitive verbs are present. Eventually, linguistic elements causing main difficulties in processing are those definite NPs which cannot be distinguished by the parser on the basis of semantic selectional restrictions, and are involved in non-local (or long-distance) dependencies. The underlying idea is that higher costs are related to the integration of new referential material which needs to be coreferential with previously mentioned antecedents: this is regarded heavier but on the same level of third person pronouns, followed by easier to integrate Proper Nouns that can be identified uniquely in the world.

Following this line of reasoning, we take complexity measures to be sensible to non-canonical structures that are pragmatically motivated and are used to encode structured meaning with high informational content, related to the FOCUS/TOPIC non-argument functions in LFG. Non-canonical structures can be said to help the reader or interlocutor to better grasp the intended (pragmatically) relevant meaning in the context of use (see Birner & Ward, 2004;2006). In Levy et al.(2012) the authors “report an investigation into the online processing of non-projective dependencies in the context of prominent contemporary theories of syntactic comprehension.”(ibid.3) which is totally dedicated to extraposed Relative Clauses (hence ERC) in order to show that readers develop a high

level of expectancies for the presence of a possible non-projective or noncanonical modifying structure of an already computed NP head. Predictability of a certain noncanonical structure (hence NCS) highly depends on its frequency of use in given contexts. Italian noncanonical structures are relatively highly represented, as the following table shows, in our treebank called VIT(Delmonte et al., 2007), where they have been explicitely marked with the labels indicated below:

NCS/ Types	LDC	S_DIS	S_TOP	S_FOC	DiscMods	Total | % Non | % NCS | Project. | / TSSe
Counts	251	1037	2165	266	12,437	16,156 | 7% | 84.59%

Table 1: Non-projective/noncanonical structures in VIT divided up by functional types.

The final percentage is computed on the total number of constituents, amounting to 230,629. [6] If we compare these data with those made available by Mambrini & Passarotti(2013) for Latin, where the same index amounts to 6.65% - data taken from the Latin Dependency Treebank containing some 55,000 tokens-, we can see that Italian and Latin are indeed very close. The second percentage is computed by dividing up number of NCS/TotalNumber of SimpleSentences. As for tree projectivity in the Penn Treebank (here marked as PT), numbers are fairly low as can be seen in the following table.

TBs/NCS	**NCS**	**UnxSubj**	**TUtt**	**TSSen**
VIT Totals	3,719	9,800	10,200	19,099
PT Totals	7,234	2,587	55,600	99,002
VIT %	27.43%	51.31%		
PT %	13.16%	0.26%		

Table 2: NonCanonicalStructures and Unexpressed Subjects in VIT and PT.

Total number of constituents for PT amounts to 720,086. Percent of NCS are computed on the number of Total Utterances, while percentage of Unexpressed Subjects are computed on the number of Total Simple Sentences. The nonprejectivity index for PT would then amount to 0.01004%. Expectancies for an Italian speaker for presence of a NCS are thus predictable to be fairly high, due to processing difficulties raised by number of Unexpressed Subjects(UnxSubj), in particular, as discussed above. This will not apply to an English speakers because NCS are unfrequent and used only in specific contexts and situations.

4 An Experiment with Rule-Based and Statistical Dependency Parsers

In this final section we will present results of the analysis of a highly non-canonical text, the fable of "the 3 little pigs"(see Appendix 1). We have been using what are regarded best parsers of Italian today, as they have been evaluated in EVALITA (Bosco & Mazzei, 2012; see www.evalita.it) evaluation campaigns. Most of them are accessible on the web[7]. Two of the parsers are fully statistical . TextPro and DeSR – while the others are hybrid rule-based/statistical parsers – TULE and VISL. We checked the output of the four parsers and marked both labeled and dependency errors. Results are shown in the table below. In Table 3. we report errors made by each parser, divided up into two classes, Labels and Dependencies. We then indicate number of words, and words with punctuation that we use to make statistics on Error Rate of each parser. As can be easily gathered, Mean values average well over 20% error rate with a resulting accuracy of less than 80%. Another important measure we report is number of fully errorless sentences: their means is below 5 over 25, that is only one fifth of the sentences are able to provide a correct mapping to semantics. The parsers share their

[6] LDC = Left Dislocated Complement S_DIS = Dislocated Subject (postposed); S_TOP = Topicalized Subject (preposed); S_FOC = Focalized Subject (inverted); DisMods = Discontinuous Modifers which include PP, PbyP, PofP, VP, RelCl, AP

[7] http://beta.visl.sdu.dk/visl/it/parsing/automatic/parse.php
http://hlt-services2.fbk.eu/textpro/?page_id=56
http://tanl.di.unipi.it/it/
http://www.tule.di.unito.it/

fully correct sentences on the total number of five which are in fact the ones containing only canonical structures. All remaining sentences have one or more noncanonical structure, thus showing that statistical data-driven parsers are unfit to cope with narrative Italian text. This is partly due to their models which are unable to generalize to noncanonical structure. In addition, their algorithms are unable to use global measures of grammaticality to improve their local choices, and in many cases cannot correct the choice of labeling as OBJect an inverted or simply dislocated NP, when the SUBJect is eventually missing, thus leaving the structure without a SUBJect.

	TULE	VISL	DeSR	TextPro	Means	StanfordLP
Label	*37*	*45*	*52*	*57*	*47,75*	*17*
Deps	*32*	*56*	*44*	*59*	*47,75*	*24*
WordsNoPu	*349*	*348*	*361*	*360*	*354,5*	*0*
totalWords	*386*	*386*	*387*	*387*	*386,5*	*395*
CorrectSent	*7 over 26*	*4 over 24*	*5 over 27*	*3 over 26*	*4,75*	*14 over 27*
ErrRtNoPu	*19,77%*	*29,02%*	*26,59%*	*32,22%*	*26,90%*	*10,38%*
ErrorRate	*17,88%*	*26,17%*	*24,81%*	*29,97%*	*24,71%*	*10,38%*

Table 3: Experiment with four parsers of Italian + Stanford LP

We then used the same text in its English translation to check what Stanford Parser was able to do. Translating Italian fables into English erases most cases of noncanonical structures so that the result is a much simpler (canonical) text to parse. As can be seen, now the number of fully correct sentence is over half the whole set 14/27 and error rate has decreased dramatically to 10% yielding a 90% accuracy.

In the Appendix, we report the Italian text, where we marked with underscores all noncanonical structures, and with italics all sequences of ambiguous attachment which may cause errors. In addition, the story has 37 Null Subjects which have been correctly found only by TULE parser - but not bound to an antecedent or syntactic controller[8]. Then there are 9 long distance dependencies, relative and wh- clauses, all correctly bound again by TULE. And 9 additional so-called open adjuncts, participials and adjectival phrases which have only been marked as verb dependent, but which also need argument dependency, since they all require agreement to be checked – this being a shortcoming of dependency structure representation. None of these structures have been marked, and in fact they constitute one of the elements causing main errors, also for Stanford parser.

5 Conclusion

Computing sentence complexity has been related to the level of F-structure mapping following LFG theoretical approach, which has no relation at all with Linear/Dominance precedence or Distance sensitive computation, these latter parameters having no consideration for lexical properties of Head-Dependence or Constituent-Governor relations. To account for the processing or parsing performance related evaluation, at first we proposed to consider the working of the mental parser to be simulated by that of an incremental ATN parser. On the contrary, Blache(2011) defines a list of six constraints on the basis of which to compute linguistic complexity, which are based solely on structural criteria. Other parameters are related to frequency of words, number of phrase-level categories in a representation, ambiguity level of each POS-tagged word, or simply number of words in the structure. Dependency-structure related criteria are then used to evaluate complexity by combining Dependency and Linearity related constraint violations. The definition of a Global Difficulty Model is then based on the sum of Local complexity evaluation indices called ID (Incomplete Dependency), DLT (Dependency Locality Theory) and Depth (depth of syntactic structure). However, this approach is applied by Blache(2011) to the output of a parser, be it dependency or constituency based, be it rule or statistically based.

[8] LFG theory distinguishes four cases of empty categories: syntactic controlled ones – so-called long distance dependencies; lexical control, subcategorized infinitivals; structural control, open adjuncts; and anaphoric control, empty pronominals.

In a more recent paper Blache(2015) presents an approach to complexity evaluation which also encompasses a parsing phase. It is based on the idea that chunking and their fusion into "constructions" are a sufficient structural level for the default identification of meaningful parts of the sentence. This is then related to "Properties" of the relations between words which are again constituted by six constraints proposed in his previous papers. Finally, the parsing process is defined as consisting in "evaluating properties when scanning a new word in the sentence"(ibid., p.14). He then distinguishes two possible parsing styles: "shallow" (when no context is needed and linearity and co-occurrence are enough to define relations between adjacent words); and "deep" (when building different partitions of the set of words in the sentence is needed and then evaluating the properties of the current word is possible). However, this second case is no longer incrementally justified, requiring a global evaluation of the sentence. So, the new approach is called "hybrid parsing" where chunks are aggregated in "buffers" – these corresponding strictly speaking to the "registers" of our ATN parser.

Kluender(1998) demonstrates the validity of a processing approach to wh- islands in which the cost of holding an uninterpreted constituent in working memory explains the increased level of complexity, using ERPs. The approach is the constructional one where there is no constituent displacement nor movement to account for, but following also Falk(2009; but see also Goldberg, 2005), the explanation is given by the multifunctionality associated to noncanonical structural elements which in LFG theory receive the nonargument roles of FOCUS/TOPIC. Constructions being typically language dependent, may vary from English to Italian, a language this latter that allows more freedom to the position of constituents in the sentence. Following this approach, explaining the additional load of FOCUS constituents in postverbal position becomes very easy: the empty slot in SUBJect position is at first associated to little_pro (the Null Subject) and then searched in preceding discourse, for a coreferential interpretation. When the FOCUS constituent in postverbal position is reached, a new interpretation is proposed which substitutes the previous one, eliminating the presence of a Null Subject in working memory. In this way, all cases of wh- FOCUS/TOPIC resemble a case of "garden path" from a processing point of view. On the contrary, those cases of NP FOCUS structures like the ones determined by presence of an unaccusative verb, or simply a focussed passive structure, have no additional processing cost. Once the verb is processed, seen that Italian constructions allow both preverbal and postverbal SUBJect, the computation goes straightforwardly: in the case of preverbal position the NP is both SUBJect and TOPIC, whereas in the case of postverbal position, the NP is SUBJect and FOCUS in full accordance with the information theory assumption that Old comes before New(see Birner & Ward, 2004;2006).

References

Birner, B. & Ward, G. 2004. Information structure and non-canonical syntax. In Horn, L. & Ward, G. The Handbook of Pragmatics. London: Blackwell. 153-174.

Birner, B. & Ward, G. 2006. Information structure. In Aarts, B. & McMahon, A. The Handbook of English Linguistics. London: Blackwell. 291-317.

Blache P. (2011). A computational model for linguistic complexity, in Proceedings of the first International Conference on Linguistics, Biology and Computer Science.

Blache P. (2015). Hybrid Parsing for Human Language Processing, in B. Sharp, W. Lubaszewski and R. Delmonte (eds) 2015. Natural Language Processing and Cognitive Science, Libreria Editrice Cafoscarina, Venice, p.9-20.

Bosco C. and A. Mazzei. 2012. The evalita 2011 parsing task: the dependency track. In Working Notes of EVALITA 2011, Rome, Italy.

Bresnan J. (ed.), 1982. The Mental Representation of Grammatical Relations, The MIT Press, Cambridge MA.

Cristiano Chesi, Il processamento in tempo reale delle frasi complesse, in E.M.Ponti, M.Budassi(eds.), 2016. Compter Parler Soigner - Tra linguistica e intelligenza artificiale, Pavia University Press, Pavia, p. 21-38. http://archivio.paviauniversitypress.it/oa/9788869520389.pdf

Delmonte R., 1984, La Syntactic Closure nella Teoria della Performance, Quaderni Patavini di Linguistica 4, 101-131.

Delmonte R., 1985. Parsing Difficulties & Phonological Processing in Italian, Proceedings of the 2nd Conference of the European Chapter of ACL, Geneva, 136-145.

Delmonte R.(2000), Parsing Preferences and Linguistic Strategies, in LDV-Forum - Zeitschrift fuer Computerlinguistik und Sprachtechnologie - "Communicating Agents", Band 17, 1,2, (ISSN 0175-1336), pp. 56-73.

Delmonte R.(2002), Relative Clause Attachment And Anaphora: Conflicts In Grammar And Parser Architectures, in A.M. Si Sciullo(ed), Grammar and Natural Language Processing, UQAM, Montreal, pp.63-87.

Delmonte R., 2004. Parsing Arguments and Adjuncts, *Proc. Interfaces Conference, IEEE - ICEIS (the International Conference on Enterprise Information Systems*), Pescara, 1-21.

Delmonte R., 2005, Deep & Shallow Linguistically Based Parsing, in A.M.Di Sciullo(ed), UG and External Systems, John Benjamins, Amsterdam/Philadelphia, pp.335-374.

Delmonte R. Bristot A., Tonelli S. (2007), VIT - Venice Italian Treebank: Syntactic and Quantitative Features, in K. De Smedt, Jan Hajic, Sandra Kübler(Eds.), Proc. Sixth International Workshop on Treebanks and Linguistic Theories, Nealt Proc. Series Vol.1, pp. 43-54.

Falk Y., 2009. Islands: A Mixed Analysis. In Butt and King 2009, *Proceedings of the LFG09 Conference*. Stanford, CA: CSLI Publications, 261–281.

Ford M., J.Bresnan, R.M.Kaplan, 1981. A competence-based theory of syntactic closure, in J.Bresnan (ed.), The Mental Representation of Grammatical Relations, MIT Press, 727-796.

Francis W. N. and H. Kucera. 1964. Brown Corpus. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Revised 1971, Revised and Amplified 1979.

Friedmann N., A. Belletti, L. Rizzi (2009). "Relativized relatives: Types of intervention in the acquisition of A-bar dependencies". Lingua, 119.1, pp. 67–88.

Hara T., Y. Miyao, J. Tsujii, 2010. Evaluating the Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. in H.Bunt et al.(eds.), Trends in Parsing Technologies, Text, Speech and Language Technology 43. 257-270.

Haug, Dag Trygve Truslew. 2012. From dependency structures to LFG representations. In Miriam Butt & Tracy Holloway King (eds.), Proceedings of the LFG12 conference, 271–291. CSLI Publications.

Gibson E. (1998). "Linguistic complexity: locality of syntactic dependencies". Cognition, 68.1, pp. 1–76.

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

Glushan Z. and A. Calabrese, 2014. Context Sensitive Unaccusative in Russian and Italian, Proceedings of the 31st West Coast Conference on Formal Linguistics, ed. Robert E. Santana-LaBarge, 207-217.

Goldberg, A.E., 2005. Constructions, Lexical Semantics and the Correspondence Principle: Accounting for Generalizations and Subregularities in the Realization of Arguments. In *The Syntax of Aspect*, Nomi Erteschik-Shir and Tova Rapoport (eds.). Oxford University Press.

Jinho D. Choi, Joel Tetreault, Amanda Stent, 2015. It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 387–396, Beijing, China.

Kaplan R.M., J. Bresnan, 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation, in J.Bresnan(ed.), The Mental Representation of Grammatical Relations, The MIT Press, Cambridge MA, pp. 173-281, republished in M.Dalrymple, R.M.Kaplan, J.T.Maxwell, A.Zaenen, (1995), Formal Issues in Lexical-Functional Grammar, CSLI, Stanford, pp- 1-102 (numbering in the paper is referred to this version).

Kluender, R. (1998) On the distinction between strong and weak islands: a processing perspective. Syntax Semantics 29, 241–279.

Levy R., Fedorenko E., Breen M., T. Gibson, 2012. The Processing of Extraposed Structures in English, Cognition, 122(1).

Mambrini F., M.Passarotti, 2013. Non-projectivity in the Ancient Greek Dependency Treebank, Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), 177–186, Prague.

Marcus, Mitchell P., Beatrice Santorini, and MaryAnn Marcinkiewicz, 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313–330.

McClosky David, Eugene Charniak, Mark Johnson, 2010. Automatic Domain Adaptation for Parsing, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 28–36.

Partee, Barbara et al. 2011. Russian Genitive of Negation Alternations: the role of verb semantics. *Scando Slavica* 57:2, 135-159.

Rizzi L., 1982. *Issues in Italian Syntax*, Oordrecht, Foris Pub.

Surdeanu Mihai, Richard Johansson, Adam Meyers, Lluıs Marquez, Joakim Nivre, 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies, in CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, 159–177.

Wanner E., M.Maratsos(1978), An ATN Approach to Comprehension, in M.Halle, J.Bresnan, G.A.Miller(eds.), 1978. Linguistic Theory and Psychological Reality, MIT Press, 119-161.

APPENDIX 1 – The Story of the Three Little Pigs

C'erano una volta tre fratelli porcellini che vivevano felici nella campagna. Nello stesso luogo però viveva anche un terribile lupo che si nutriva proprio di porcellini grassi e teneri. Questi allora, per proteggersi dal lupo, decisero di costruirsi ciascuno una casetta. Il maggiore, Jimmy che era saggio, lavorava di buona lena e costruì *la sua casetta con* solidi mattoni *e cemento*. Gli altri, Timmy e Tommy, pigri se la sbrigarono in fretta costruendo *le loro casette con* la paglia *e con pezzetti di legno*. I due porcellini pigri passavano le loro giornate suonando e cantando una canzone che diceva: chi ha paura del lupo cattivo. Ma ecco che improvvisamente il lupo apparve alle loro spalle. Aiuto, aiuto, gridarono i due porcellini e cominciarono a correre più veloci che potevano verso la loro casetta per sfuggire al terribile lupo. Questo intanto si leccava già i baffi pensando al suo prossimo pasto così invitante e saporito. Finalmente i porcellini riuscirono a raggiungere la loro casetta e vi si chiusero dentro sbarrando la porta. Dalla finestra cominciarono a deridere il lupo cantando la solita canzoncina: chi ha paura del lupo cattivo. Il lupo stava intanto pensando al modo di penetrare nella casa. Esso si mise ad osservare attentamente la casetta e notò che non era davvero molto solida. Soffiò con forza un paio di volte e la casetta si sfasciò completamente. Spaventatissimi i due porcellini corsero *a perdifiato verso* la casetta del fratello. "Presto, fratellino, aprici! Abbiamo *il lupo alle* calcagna". Fecero appena in tempo ad entrare e tirare il chiavistello. Il lupo *stava già arrivando deciso* a non rinunciare al suo pranzetto. Sicuro di abbattere anche la casetta di mattoni il lupo si riempì *i polmoni di* aria e cominciò a soffiare *con forza alcune volte*. Non c'era niente da fare. La casa non si mosse di un solo palmo. Alla fine esausto il lupo si accasciò a terra. I tre porcellini si sentivano *al sicuro nella* solida casetta di mattoni. Riconoscenti i due porcellini oziosi promisero *al fratello che* da quel giorno anche essi avrebbero lavorato sodo.[9]

Once upon a time there were three little pigs who lived happily in the countryside. But in the same place lived a wicked wolf who fed precisely on plump and tender pigs. The little pigs therefore decided to build a small house each, to protect themselves from the wolf. The oldest one, Jimmy who was wise, worked hard and built his house with solid bricks and cement. The other two, Timmy and Tommy, who were lazy settled the matter hastily and built their houses with straw and pieces of wood. The lazy pigs spent their days playing and singing a song that said, "Who is afraid of the big bad wolf?" And one day, lo and behold, the wolf appeared suddenly behind their backs. "Help! Help!", shouted the pigs and started running as fast as they could to escape the terrible wolf. He was already licking his lips thinking of such an inviting and tasty meal. The little pigs eventually managed to reach their small house and shut themselves in, barring the door. They started mocking the wolf from the window singing the same song, "Who is afraid of the big bad wolf?" In the meantime the wolf was thinking a way of getting into the house. He began to observe the house very carefully and noticed it was not very solid. He huffed and puffed a couple of times and the house fell down completely. Frightened out of their wits,the two little pigs ran at breakneck speed towards their brother's house. "Fast, brother, open the door! The wolf is chasing us!" They got in just in time and pulled the bolt. Within seconds the wolf was arriving, determined not to give up his meal. Convinced that he could also blow the little brick house down, he filled his lungs with air and huffed and puffed a few times. There was nothing he could do. The house didn't move an inch. In the end he was so exhausted that he fell to the ground. The three little pigs felt safe inside the solid brick house. Grateful to their brother, the two lazy pigs promised him that from that day on they too would work hard.[i]

[9] This is a modified version of the Italian version of the original story, which I have done to make it shorter.