Searching treebanks for functional constraints: cross-lingual experiments in grammatical relation assignment

[1,3]Felice Dell'Orletta, [2]Alessandro Lenci, [3]Simonetta Montemagni, [3]Vito Pirrelli

[1]Università di Pisa, Dipartimento di Informatica
Largo B. Pontecorvo 3, 56100 Pisa, Italy
felice.dellorletta@ilc.cnr.it

[2]Università di Pisa, Dipartimento di Linguistica
Via Santa Maria 36, 56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

[3]Istituto di Linguistica Computazionale (ILC) - CNR
Area della Ricerca - via G. Moruzzi 1, 56100 Pisa, Italy
{simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Abstract

We report here on a detailed quantitative analysis of distributional language data of both Italian and Czech, highlighting the relative contribution of a number of distributed grammatical factors to sentence-based identification of subjects and direct objects. The work is based on a Maximum Entropy model of stochastic resolution of grammatical conflicting constraints, and is demonstrably capable of putting explanatory theoretical accounts to the challenging test of an extensive, usage-based empirical verification.

1. Introduction

Treebanks allow for multiple uses: by linguists, which may search for examples (or counter-examples) for a given theory or hypothesis; by psycholinguists, interested in computing frequencies and comparing them with human preferences; by computational linguists, for tasks such as lexicon and grammar induction or parser evaluation. New challenging questions involve the use of Treebanks to determine the typology of factors playing a role in specific natural language learning and processing tasks as well as their relative salience. Answers to these questions can shed novel light on both genuinely linguistic and psycholinguistic issues as well as usefully be exploited for parsing purposes. In the present paper we intend to illustrate extensive use of Treebanks for the discovery and comparative assessment of typologically relevant and linguistically motivated constraints on cross-linguistic parsing issues. For these purposes we shall focus on a detailed evaluation of corpus-based discovery procedures of this kind applied to interestingly different languages such as Italian and Czech.

Current research in natural language learning and processing supports the view that grammatical competence consists in mastering and integrating multiple, parallel constraints (Seidenberg and MacDonald 1999, MacWhinney 2004). Moreover, growing consensus exists on two major properties of grammatical constraints, i.e. i.) that they are probabilistic "soft constraints" (Bresnan *et al.* 2001), and ii.) that they have an inherently functional nature, involving different types of linguistic (and non linguistic) information (syntactic, semantic, etc.). These features emerge clearly when we focus on one of the core aspects of grammatical competence: the ability to properly identify *syntactic relations*. Psycholinguistic evidence shows that speakers learn to identify sentence subjects and direct objects by combining various types of probabilistic, functional cues, such as word order, noun animacy, definiteness, agreement, etc. An important observation is that the relative prominence of each of these cues can considerably vary cross-linguistically. Bates *et al.* (1984), for example, argue that while, in English, word order is the most effective cue for subject-object identification (henceforth *SOI*) both in syntactic processing and during the child's syntactic development, the same cue plays second fiddle in languages such as Italian or German.

If grammatical constraints are inherently probabilistic (Manning 2003), the path through which adult grammar competence is acquired can be viewed as the process of building a stochastic model out of the linguistic input. In computational linguistics, *Maximum Entropy* (henceforth MaxEnt) models have proven to be robust statistical learning algorithms that perform well in a number of processing tasks. In this paper, we illustrate an application of the MaxEnt model to the processing of subjects and direct objects in Italian and Czech.

2. Subjects and objects in Czech and Italian

Grammatical relations - such as subject (*S*) and direct object (*O*) - can be variously encoded in languages, the two most widespread strategies being: i) structural encoding through *word order*, and ii) morpho-syntactic marking. In turn, morpho-syntactic marking can apply either on the noun head only, in the form of *case inflections*, or on both the noun and the verb, in the form of agreement marking. (Croft 2003). Besides formal coding, the distribution of subjects and object is also governed by semantic and pragmatic factors, such as noun animacy, definiteness, topicality, etc. As a result, there exists a variety of linguistic clues jointly co-operating in making a particular noun phrase as the subject or direct object of a sentence. Crucially for our present purposes, cross-linguistic variation does not only concern the particular strategy used to encode *S* and *O*, but also the *relative strength* that each factor plays in a given language. For instance, while English word order is by and large the dominant clue to identify *S* and *O*, in other languages the presence of a rich morphological system allows word order to have a much looser connection with the coding of grammatical relations, thus playing a secondary role in their identification. Moreover, there are languages where semantic and pragmatic constraints such as animacy and/or definiteness play a predominant role in the assignment of grammatical relations. Still, a large spectrum of variations exists, ranging from languages where *S must* have a higher degree of animacy and/or definiteness relative to *O*, to languages where this constraint only takes the form of a softer statistical preference (cf. Bresnan *et al.* 2001).

The goal of this paper is to explore the area of this complex space of variation through careful assessment of the distribution of *S* and *O* in Italian and Czech. For our present analysis, we have used a MaxEnt statistical model trained on data extracted from two syntactically annotated corpora: the *Prague Dependency Treebank* (PDT, Bohmova *et al.* 2003) for Czech, and the *Italian Syntactic Semantic Treebank* (ISST, Montemagni *et al.* 2003) for Italian. These corpora have been chosen not only because they are the largest syntactically annotated resources for the two languages, but also because of their high degree of comparability, since they both adopt a dependency-based annotation scheme.

Czech and Italian provide a very interesting vantage point for the cross-lingual analysis of grammatical variation. They are both Indo-European languages, but they do not belong to the same family: Czech is a West Slavonic language, while Italian is a Romance language. There are two major features they share: i) the free order of grammatical relations with respect to the verb; ii) the possible absence of an overt subject. Nevertheless, they also greatly differ because of the virtual non-existence of case marking in Italian (with the only marginal exception of personal pronouns), and for the degree of word order freedom in the two languages. Empirical evidence supporting the latter claim is provided in Table 1, which reports data extracted from PDT and ISST. Notice that although in both languages *S* and *O* can occur either pre-verbally or post-verbally, Czech and Italian greatly differ in their propensity to depart from the SVO order. While in Italian preverbal *O* is highly marked (1.90%), in Czech more than the 30% of *O* occur before the verb. The situation is different for *S*, which occurs after the verb in Italian in 22, 21% of the cases. In Czech, *S* also shows a higher tendency (40%) to occur post-verbally. For sure, one can argue that, in spoken Italian, the number of pre-verbal objects is actually higher, because of the greater number of left dislocations and topicalizations. However, this fact can not be used to explain away the differences in distribution between the two corpora, since both PDT and ISST contain written language only. We thus suggest that there is clear empirical evidence in favour of a systematic, higher word-order freedom in Czech, arguably related to the well-known fact that in Czech the position of a lexical unit in the sentence is governed primarily by the information structure of the latter, with the element carrying new information showing a tendency to occur sentence-finally (Stone 1990). For our present concerns, however, aspects of information structure, albeit central in Czech grammar, could not possibly be taken into account, as they were not marked-up in the Italian corpus.

According to the data reported in Table 1, Czech and Italian obey to similar patterns with respect to the relationship between animacy and grammatical relations. *S* and *O* in ISST were automatically annotated for animacy using the SIMPLE Italian computational lexicon (Lenci *et al.* 2000) as a background semantic resource. The annotation was then checked manually. Czech *S* and *O* were annotated for animacy using Czech WordNet (Pala and Smrz 2004); it is worth remarking that in Czech animacy annotation was done only automatically, without any manual revision. In Italian, there is a strong asymmetry in the distribution of animate nouns in subject and object roles: over 50% of ISST subjects are animate, while only 10% of the objects are animate. Such a trend is also confirmed in Czech – although to a lesser extent - with 34.10% of animate subjects vs. 15.42% of objects.[1] Such an overwhelming preference for animate subjects in corpus data suggests that animacy plays a very important role for *S* and *O* identification in both languages.

Corpus data also provide interesting evidence concerning the actual role of morpho-syntactic constraints in the distribution of grammatical relations. *Prima facie*, agreement and case represent the strongest and most directly accessible clues to decide about *S/O* in a sentence, as they are marked both overtly and locally. This is also confirmed by psycholinguistic data, showing that subjects prefer to rely on these clues to identify *S/O*. However, such formal clues are not always available in context. In fact, agreement represents conclusive evidence for *SOI* only when a nominal constituent and a verb do not agree in number and/or person (as in *leggono il libro* '(they) read the book'). When N and V share the same person and number the impact of agreement for grammatical relation assignment is neutralised, as in *il bambino legge il libro* 'the child reads the book' or in *ha dichiarato il presidente* 'the president declared'. It is interesting to note that in ISST more than 58% of *O* agree with their governing V, thereby being formally undistinguishable from *S* on agreement features. PDT also exhibits a similar ratio, with 56% of *O* agreeing with their verb. Analogous considerations apply to case marking, whose perceptual reliability is undermined by morphological syncretism, when different cases are realized through the same marker. Czech data reveal the massive extent of this phenomenon and its impact on *SOI*. As reported in Table 2, more than 56% of *O* extracted from PDT are formally undistinguishable from *S* in their case ending. Similarly, 45 % of *S* are formally undistinguishable from *O* on the same ground. All in all, this means that in 50% of the cases a Czech noun can not be understood as *S/O* of a sentence by relying on overt case marking only.

To sum up, corpus data lend support to the idea that in both Italian and in Czech *SOI* is governed by a complex interplay of probabilistic constraints of a different nature: morpho-syntactic, semantic, word order etc. Moreover, distributional asymmetries in language data seem to provide a fairly reliable statistical basis upon which relevant probabilistic constraints can be bootstrapped and combined consistently, in order to model their different degrees of salience in the two languages. The following section illustrates how a MaxEnt model can be used to bootstrap constraints and their interaction from language data.

[1] In fact, the considerable difference in animacy distribution between the two languages might only be an artefact of the way we annotated Czech nouns semantically, on the basis of their context-free classification in the Czech WordNet.

		Czech		Italian	
		Subj	Obj	Subj	Obj
Pos	Pre	59.82%	30.27%	77.79%	1.90%
	Post	40.18%	69.73%	22.21%	98.10%
	All	100.00%	100.00%	100.00%	100.00%
Agr	Agr	98.50%	56.54%	97.73%	58.33%
	NoAgr	1.50%	43.46%	2.27%	41.67%
	All	100.00%	100.00%	100.00%	100.00%
Anim	Anim	34.10%	15.42%	50.18%	10.67%
	NoAnim	65.90%	84.58%	49.82%	89.33%
	All	100.00%	100.00%	100.00%	100.00%

Table 1 –*Distribution of Czech and Italian S and O wrt word order, agreement and noun animacy*

	Czech	
	Subj	Obj
Nominative	53.83%	0.65%
Accusative	0.15%	28.30%
Dative	0.16%	9.54%
Genitive	0.22%	2.03%
Instrumental	0.01%	3.40%
Ambiguous	45.63%	56.08%
All	100.00%	100.00%

Table 2 - *Distribution of Czech S and O wrt case*

3. Maximum Entropy modeling

The Maximum Entropy (ME) framework offers a mathematically sound way to build a probabilistic model for SOI, which combines different linguistic cues. Given a linguistic context *c* and an outcome $a \in A$ that depends on *c*, in the ME framework the conditional probability distribution $p(a|c)$ is estimated on the basis of the assumption that no *a priori* constraints must be met other than those related to a set of features $f_j(a,c)$ of *c*, whose distribution is derived from the training data. It can be proven that the probability distribution *p* satisfying the above assumption is the one with the highest entropy, is unique and has the following exponential form (Berger *et al.* 1996):

$$(1) \qquad p(a \mid c) = \frac{1}{Z(c)} \prod_{j=1}^{k} \alpha_j^{f_j(a,c)}$$

where $Z(c)$ is a normalization factor, $f_j(a,c)$ are the values of *k* features of the pair (*a*,*c*) and correspond to the linguistic cues of *c* that are relevant to predict the outcome *a*. Features are extracted from the training data and define the constraints that the probabilistic model *p* must satisfy. The parameters of the distribution $\alpha_1, ..., \alpha_k$ correspond to *weights* associated with the features, and determine the relevance of each feature in the overall model. In the experiments reported below feature weights have been estimated with the Generative Iterative Scaling (GIS) algorithm implemented in the AMIS software (Miyao and Tsujii 2002).

We model *SOI* as the task of predicting the correct syntactic function $\varphi \in$ {*subject*, *object*} of a noun occurring in a given syntactic context σ. This is equivalent to build the conditional probability distribution $p(\varphi|\sigma)$ of having a syntactic function φ in a syntactic context σ. Adopting the ME approach, the distribution *p* can be rewritten in the parametric form of (1), with features corresponding to the linguistic contextual cues relevant to *SOI*. The context σ is a pair $<v_\sigma, n_\sigma>$, where v_σ is the verbal head and n_σ its nominal dependent in σ. This notion of σ departs from more traditional ways of describing an *SOI* context as a *triple* of one verb and two nouns in a certain syntactic configuration (e.g, *SOV* or *VOS*, etc.). In fact, we assume that *SOI* can be stated in terms of the more local task of establishing the grammatical function of a noun *n* observed in a verb-noun pair. This simplifying assumption is consistent with the claim in MacWhinney *et al.* (1984) that *SVO* word order is actually derivative from *SV* and *VO* local patterns and downplays the role of the transitive complex construction in sentence processing. Evidence in favour of this hypothesis also comes from corpus data: for instance, in ISST complete subject-verb-object configurations represent only 26% of the cases, a small percentage if compared to the 74% of verb tokens appearing with either a subject or an object only. Due to the comparative sparseness of canonical *SVO* constructions in Italian, it seems more reasonable to assume that children should pay a great deal of attention to both *SV* and *VO* units as cues in sentence perception (Matthews *et al.* in press). Reconstruction of the whole lexical *SVO* pattern can accordingly be seen as the end point of an acquisition process whereby smaller units are re-analyzed as being part of more comprehensive constructions. This hypothesis is more in line with a *distributed* view of canonical constructions as derivative of more basic local positional patterns, working together to yield more complex and abstract constructions. Last but not least, assuming verb-noun pairs as the relevant context for *SOI* allows us to simultaneously model the interaction of word order variation with pro-drop.

4. Feature selection

The most important part of any MaxEnt model is the selection of the context features whose weights are to be estimated from data distributions. Our feature selection strategy is grounded on the main assumption that *features should correspond to theoretically and typologically well-motivated contextual cues*. This allows us to evaluate the probabilistic model also with respect to its consistency with current linguistic generalizations. In turn, the model can be used as a probe into the correspondence between theoretically motivated generalizations and usage-based empirical evidence.

Features are binary functions $f_{k_i,\varphi}$ (φ,σ), which test whether a certain *cue* k_i for the function φ occurs in the context σ. For our MaxEnt model, we have selected different features types that test *morpho-syntactic*, *syntactic*, and *semantic* key dimensions in determining the distribution of *S* and *O*.

Morpho-syntactic features. These include *N-V agreement,* for Italian and Czech, and *case*, only for Czech. The combined use of such features allow us not only to test the impact of morpho-syntactic information on *SOI*, but also to analyze patterns of cross-lingual variation stemming from language specific morphological differences, e.g. lack of case marking in Italian.

Word order. This feature essentially test the position of the noun wrt the verb, for instance:

$$(2)\ f_{post,subj}(subj,\sigma) = \begin{cases} 1 & \textit{if noun}_{\sigma}\textit{.pos = post} \\ 0 & \textit{otherwise} \end{cases}$$

Animacy. This is the main semantic feature, which tests whether the noun in *σ* is *animate* or *inanimate* (cf. §.2). The centrality of this cue for grammatical relation assignment is widely supported by typological evidence (cf. Aissen 2003, Croft 2003). The Animacy Markedness Hierarchy - representing the relative markedness of the associations between grammatical functions and animacy degrees – is actually assigned the role of a functional universal principle in grammar. The hierarchy is reported below, with each item in these scale been less marked than the elements to its right:

Animacy Markedness Hierarchy
Subj/Human > Subj/Animate > Subj/Inanimate
Obj/Inanimate > Obj/Animate > Obj/Human

Markedness hierarchies have also been interpreted as probabilistic constraints estimated from corpus data (Bresnan *et al.* 2001). In our MaxEnt model we have used a reduced version of the animacy markedness hierarchy in which human and animate nouns have been both subsumed under the general class *animate*.

Definiteness tests the degree of "referentiality" of the noun in a context pair *σ*. Like for animacy, definiteness has been claimed to be associated with grammatical functions, giving rise to the following universal markedness hierarchy Aissen (2003):

Definiteness Markedness Hierarchy
Subj/Pro > Subj/Name > Subj/Def > Subj/Indef
Obj/Indef > Obj/Def > Obj/Name > Obj/Pro

According to this hierarchy, subjects with a low degree of definiteness are more marked than subjects with a high degree of definiteness (for objects the reverse pattern holds). Given the importance assigned to the definiteness markedness hierarchy in current linguistic research, we have included the definiteness cue in the MaxEnt model. In our experiments, for Italian we have used a "compact" version of the definiteness scale: the definiteness cue tests whether the noun in the context pair i) is a name or a pronoun ii) has a definite article iii), has an indefinite article or iv) is a "bare" noun (i.e. with no article). It is worth saying that "bare" nouns are usually placed at the bottom end of the definiteness scale. Since in Czech there is no article, we only make a distinction between proper names and not proper names.

5. Testing feature configurations for SOI

The ME model for Italian SOI has been trained on 14,643 verb-subject/object pairs extracted from ISST. For Czech SOI we used a training corpus of 37,947 verb-subject/object pairs extracted from PTD. In both cases, the training set was obtained by extracting all verb-subject and verb-object dependencies headed by an active verb and by excluding all cases where the position of the nominal constituent was grammatically constrained (e.g. clitic objects, relative clauses). It is interesting to note that in both training sets the proportion of subjects and objects relations is nearly the same: 63.06%-65.93% verb-subject pairs and 36.94%-34.07% verb-object pairs for Italian and Czech respectively.

Two different feature configurations have been used for training:

- non-lexical feature configuration (*NLC*), including only general features acting as global constraints: namely verb agreement, case (for Czech only), word order, noun animacy and noun definiteness;
- lexical feature configuration (*LC*), including verb agreement, case (for Czech only), word order, noun animacy and definiteness, and information about the verb head.

The test corpus consists of a set of verb-noun pairs randomly extracted from the reference Treebanks: 1,000 pairs for Italian and 1,373 for Czech. For Italian, 559 pairs contained a subject and 441 contained an object; for Czech, 905 pairs contained a subject and 468 an object.

The model was evaluated for both languages by calculating the percentage of correctly assigned relations over the total number of test pairs (accuracy). As our model always assigns one syntactic relation to each test pair, accuracy equals both standard precision and recall. Finally, we have assumed a baseline score of 56% for Italian and of 66% for Czech, corresponding to the result yielded by a dumb model assigning to each test pair the most frequent relation in the training corpus, i.e. subject.

5.1. Non-lexical feature configuration

Our first experiments were carried out with *NLC*. The accuracy achieved by the model on the test corpus is 89.80% for Italian and 89.22% for Czech. A more detailed analysis of errors for the two languages is reported in Table 3, showing that in Czech most errors affect the object relation (i.e. 92.57%), whereas the reverse holds for Italian, where subject identification appears to be most problematic (i.e. 79.41% of errors are subjects mistaken as direct objects). It is also interesting to note how and to what extent individual features contribute to errors. In Czech it appears that the prototypically mistaken objects are post-verbal (66.22%), inanimate (87.84%), ambiguously case-marked (91.22%) and agreeing with the verb (91.89%); the reported percentages refer to the whole error set. In Italian, mistaken subjects can be described as follows: they all occur in post-verbal position and are mostly (92.52%) inanimate. Interestingly, in either languages, the highest number of errors occurs in those cases in which *N* has the least prototypical morphosyntactic, syntactic and semantic properties for *O* (or *S*). This shows that MaxEnt has actually been able to form a precise model of the core linguistic properties that *S* and *O* have in Italian and in Czech.

A further way to evaluate the goodness of the model is by inspecting the weights associated with feature values for the two languages. They are reported in Table 4 where grey cells highlight the preference of each feature value for either subject or object identification. In both languages agreement with the verb strongly relates with the subject relation. For Czech, nominative case is strongly associated with subjects and the other cases with objects. Moreover, in both languages: preverbal subjects are strongly preferred over preverbal objects; animate

	Czech		Italian	
	Subj	Obj	Subj	Obj
Preverb	5	39	0	8
Postverb	6	98	81	13
Anim	1	7	6	13
Inanim	10	130	75	8
Nomin	0	2	Na	
Genitive	1	0		
Dative	4	0		
Accus	0	0		
Instrum	0	0		
Ambig	6	135		
Agr	4	136	67	18
NoAgr	6	1	9	2
NAAgr	1	0	5	1

Table 3 – *Typology of erros in NLC for Czech and Italian*

	Czech		Italian	
	Subj	Obj	Subj	Obj
Preverb	1.24E+00	5.40E-01	1.31E+00	2.11E-02
Postverb	8.77E-01	1.17E+00	5.39E-01	1.38E+00
Anim	1.16E+00	6.63E-01	1.28E+00	3.17E-01
Inanim	1.03E+00	9.63E-01	8.16E-01	1.23E+00
PronName	1.13E+00	7.72E-01	1.13E+00	8.05E-01
DefArt	1.05E+00	9.31E-01	1.01E+00	1.02E+00
IndefArt			6.82E-01	1.26E+00
NoArticle			9.91E-01	1.02E+00
Nomin	1.23E+00	2.22E-02	Na	
Genitive	2.94E-01	1.51E+00		
Dative	2.85E-02	1.49E+00		
Accus	8.06E-03	1.39E+00		
Instrum	3.80E-03	1.39E+00		
Agr	1.18E+00	6.67E-01	1.28E+00	4.67E-01
NoAgr	7.71E-02	1.50E+00	1.52E-01	1.58E+00
NAAgr	3.75E-01	1.53E+00	2.61E-01	1.84E+00

Table 4 - *Feature value weights in NLC for Czech and Italian*

subjects are preferred over animate objects; pronouns and proper names are typically subjects.

Let us now try to relate these feature values with the Markedness Hierarchies reported in § 4. Interestingly, for Italian, if we rank the *Anim* and *Inanim* values for subjects and objects, we observe that they distribute consistently with the *Animacy Markedness Hierarchy*: *Subj*/*Anim* > *Subj*/*Inanim* and *Obj*/*Inanim* > *Obj*/*Anim*. This is confirmed by the Czech results. Similarly, by ranking the Italian values for the definiteness features in the *Subj* column by decreasing weight values we obtain the following ordering: *PronName* > *DefArt* > *IndefArt* > *NoArt*, which nicely fits in with the *Definiteness Markedness Hierarchy* in § 4.

The so-called "markedness reversal" is replicated with a good degree of approximation if we focus on the values for the same features in the *Obj* column: the *PronName* feature represents the most marked option, followed by *IndefArt*, *DefArt* and *NoArt* (the latter two showing the same feature value). The exception here is represented by the relative ordering of *IndefArt* and *DefArt* which however show very close values. The same seems to hold for Czech, where the feature ordering for *Subj* is *PronName* > *DefArt*/*IndefArt*/*NoArt* and the reverse is observed for *Obj*. Evaluating feature salience.

The relative salience of the different constraints acting on *SOI* can be inferred by comparing the weights associated with individual feature values. For instance, Goldwater and Johnson (2003) show that ME can be successfully applied to learn constraint rankings in Optimality Theory, by assuming the parameter weights <α1, …, αk> as the ranking values of the constraints. Figure 1 shows the constraint weights ranking for the two languages; note that only positional and animacy constraints are included in the graph. The rankings in Figure 1 can be used to derive the relative salience of each constraint in Czech and Italian. Lower ranked constraints correspond to more marked syntactic configurations that are then disfavoured in *SOI*. For instance, in Italian the two animacy constraints *AnimObj* and *AnimSubj* are respectively placed near the bottom and the top end of the scale. Notwithstanding the low position of *PostSubj*, animacy is thus able to override the word order constraint and to produce a strong bias towards understanding animate nouns as subjects, even when they appear in post-verbal position. The constraint ranking thus confirms the interplay between animacy and word order in Italian, with the former playing a decisive role in assigning the syntactic function of post-verbal nouns.

5.2. Lexical feature configuration

In this experiment, the general features reported in Table 4 are integrated with verb-specific features, as illustrated below for the Italian verb *dire* 'say':

*dire*_animSog	1.228213e+00
*dire*_noanimSog	7.028484e-01
*dire*_animOgg	3.645964e-01
*dire*_noanimOgg	1.321887e+00

showing a strong preference of the verb for taking animate subjects and inanimate objects. For Italian, verb-specific features are 4,316 and for Czech 8,248. The results achieved with *LC* on the test corpora for the two languages show a significant improvement with respect to those obtained with *NLC*: accuracy is now 95.48% for Czech and 97.4% for Italian, with a significant improvement in both cases (+6.26% and +7,6% respectively).

6. Conclusions

Nowadays, probabilistic language models, machine language learning algorithms and linguistic theorizing all appear to provide substantially converging evidence supporting a view of language understanding as a process of dynamic, on-line resolution of conflicting grammatical constraints. We begin to gain considerable insights into the complex process of bootstrapping the nature and behaviour of these constraints upon observing the actual distribution of constraint configurations in perceptually salient contexts. In our view of things, this solid scientific trend not only outlines a promising framework providing fresh support to usage-based models of language acquisition through mathematical and computational simulations. It also allows us to shed light on patterns of cross-linguistic typological variation that crucially depend on the appropriate setting of model parameters. Moreover, it promises to solve, on a principled basis, traditional performance-oriented *cruces* of grammar theorizing such as degrees of human acceptability of ill-formed grammatical constructions (Hayes 2000) and the inherently graded compositionality of linguistic constructions such as morpheme-based words and word-based phrases (Bybee 2002, Hay and Baayen 2005). The work reported in the present paper is still fairly preliminary and is mainly intended to show the enormous potential of such a methodological convergence. Nonetheless, it allows us to argue that the current availability of comparable, richly annotated corpora and of mathematical tools and models for corpus exploration make time ripe for probing the space of grammatical variation, both intra- and inter-linguistically, on unprecedented levels of sophistication and granularity. All in all, we anticipate that such a convergence is likely to have a twofold impact: on the one hand, it will shed novel light on the integration of performance and competence factors in language study; on the other hand, it will make mathematical models of language increasingly able to accommodate richer and richer language evidence, thus putting explanatory theoretical accounts to the challenging test of an extensive, usage-based empirical verification.

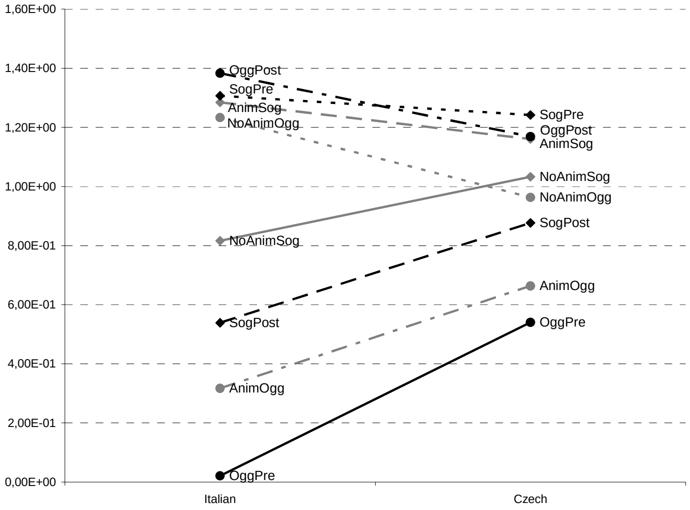

Figure 1 – *Relative ranking of animacy and positional constraints in Czech and Italian*

7. References

Bates E., MacWhinney B., Caselli C., Devescovi A., Natale F., Venza V., (1984). A crosslinguistic study of the development of sentence interpretation strategies. *Child Development*, 55: 341-354.

Bohmova A., Hajic J., Hajicova E., Hladka B., 2003. The Prague Dependency Treebank: Three-Level Annotation Scenario, in A. Abeille (ed.) *Treebanks: Building and Using Syntactically Annotated Corpora,* Kluwer Academic Publishers, pp. 103-128.

Bybee, J.,.(2002), Sequentiality as the basis of constituent structure. in T. Givón and B. Malle (eds.) *The Evolution of Language out of Pre-Language*, Amsterdam: John Benjamins. 107-132.

Croft, W (2003), *Typology and Universals. Second Edition*, Cambridge University Press, Cambridge.

Bresnan J., Dingare D., Manning C. D., (2001). Soft constraints mirror hard constraints: voice and person in English and Lummi. *Proceedings of the LFG01 Conference*, Hong Kong: 13-32.

Hay, J., R.H. Baayen, (2005), Shifting paradigms: gradient structure in morphology, *Trends in Cognitive Sciences*, 9(7): 342-348.

Hayes, B., (2000), Gradient Well-Formedness in Optimality Theory, in Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer (eds.) *Optimality Theory: Phonology, Syntax, and Acquisition*, Oxford University Press, pp. 88-120.

Lenci A. *et al.*, 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13 (4): 249-263.

MacWhinney B., (2004). A unified model of language acquisition. In J. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, Oxford University Press, Oxford.

Manning C. D., (2003). Probabilistic syntax. In R. Bod, J. Hay, S. Jannedy (eds), *Probabilistic Linguistics*, MIT Press, Cambridge MA: 289-341.

Miyao Y., Tsujii J., 2002. Maximum entropy estimation for feature forests. *Proc. HLT2002*.

Montemagni S. *et al.* 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) *Treebanks. Building and Using Parsed Corpora*, Kluwer, Dordrecht: 189-210.

Pala K., Smrz P., 2004. Building Czech Wordnet, *Romanian Journal Of Information Science And Technology*, Volume 7, Numbers 1/2, pp. 79-88.

Seidenberg M. S., MacDonald M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23(4): 569-588.

Stone, G. (1990), Czech and Slovak, in Comrie B. (ed.), *The World's Major Languages*, New York, Oxford University Press.