Corpus-based modelling of grammar variation

Felice Dell'Orletta, Alessandro Lenci,
Simonetta Montemagni, Vito Pirrelli

1. Introduction

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, Burzio 2005). During language comprehension, constraints are defined as a set of cues through which a speaker can successfully map a complex linguistic unit (a word form, a phrase, an utterance etc.) to its intended function in discourse. For example, in the Italian sentence *arriva il treno* ('the train is approaching'), the noun phrase *il treno* is understood to play the role of subject due to an integrated cluster of cues concerning the position of the subject relative to the verb, noun-verb agreement, compliance of predicate selectional restrictions and general knowledge about the world. Conversely, in language production, constraints consist in discourse functions that are jointly mapped onto linguistic forms. If we want to translate *arriva il treno* into English, we have to be aware of the peculiar functional constraint positioning the subject before the verb in the target language.

1.1. Nature of constraints

The strength of a cue as a mapping constraint within a language can vary depending on the context. In Italian, subjects are usually overtly realized before the verb, whereas the post-verbal position is typically taken by a verb complement such as a direct or an indirect object. However strong this constraint can be in Italian, it is superseded, as we just saw, in a context like *arriva il treno*, where the Verb-Subject (VS) order is by far the unmarked one, especially in presentative contexts. These facts of Italian grammar can be interpreted as evidence in support of some peculiar features of language constraints, i.e. the fact that they are probabilistic, violable, and inherently in competition with each other. This means primarily that many cues are simultaneously activated as operative constraints when an input sentence is made available to a speaker. Some of them can make contrasting predictions. In *arriva il treno*, the positioning of the noun phrase relative to the verb supports a default interpretation of

In: Andrea Sansò (ed.), *Language Resources and Linguistic Theory*, 38-55. [Materiali Linguistici 59.] Milano: Franco Angeli, 2007.

il treno as a direct object; conversely, other constraints activated by the unacccusative nature of ARRIVARE and by the informational structure of the sentence point to a subject interpretation. The speaker's processing system is asked to select among contrasting predictions on the basis of their relative cue strength, to converge onto the contextually more appropriate interpretation.

This is an important point that must be spelled out with some care. Individually, cue strength is interpreted probabilistically, as a function of the frequency with which a certain cue correlates with a certain functional interpretation. The more often an object noun phrase is seen after its governing verb, the stronger its support to an object interpretation when other novel sentences are input to the processing system of a speaker. As each context may activate more constraints, their relative contribution must be evaluated interactively, as a function of cue integration. Hence weak active constraints can co-operate to overturn the interpretation supported by an individually stronger constraint as an effect of the cumulative summation over their individual strength. It then turns out that strong constraints can be violated in concrete contexts, due to the overall pressure of a bunch of weaker constraints, coalescing to make a contrasting prediction.

1.2. Constraints and typological variation

Another important related point, which will be developed in some detail in the remainder of this paper, is that the relative prominence of a cue as a mapping constraint 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 relatively free phrase-order languages such as Italian or German. An interesting piece of distributional evidence pointing in the same direction comes from corpus data. Table 1 shows the distribution of subjects (Subj) and direct objects (Obj) in pre- (Pre) or post-verbal (Post) position in Czech and Italian, as witnessed by two comparable annotated corpora. Note that the proportion of subjects that appear after their governing verb in Czech doubles the corresponding proportion in Italian. Even more strikingly, a 30.27% of Czech preverbal objects is in sharp contrast with a very low 1.9% of preverbal objects in several thousands of sentences of Italian newspapers.

	Czech		Italian	
	Subj	Obj	Subj	Obj
Pre	59.82%	30.27%	77.79%	1.90%
Post	40.18%	69.73%	22.21%	98.10%

Table 1. *Distribution of Czech and Italian Subj and Obj wrt word order.*

The evidence provided by Table 1 is worth pausing. In a cross-linguistic perspective, this data may support two alternative interpretations. We can hypothesize that the two languages in question enforce two radically different rankings on the relative strength of the constraints governing the positioning of object noun phrases with respect to other types of constraint. According to this hypothesis, Czech ordering constraints on the surface realization of object noun phrases are ranked lower relative to other morpho-syntactic, semantic or pragmatic constraints. In Italian, the reverse ranking obtains, with ordering constraints being dominant. Such a relative dominance of ordering constraints is compatible with the proportions of post-verbal subjects in the two languages, showing a similar (albeit less strong) bias in favour of ordering constraints on subject positions in Italian (see Table 1).

In fact, the same data are compatible with a different interpretation: the relative ranking of constraint strengths is invariant in the two languages, but small differences in absolute strength values make room for larger surface differences due to constraint interaction. In other words, large differences in the distribution of subjects and objects in Czech and Italian can be accounted for in terms of the dynamic interplay of identically-ranked but differently-valued constraints.

In our view, this second hypothesis is potentially more interesting. First, it accounts for surface differences between Czech and Italian through constraint interaction, thus leaving the overall constraint ranking invariant. This move provides a more constrained and explanatory account of surface cross-lingual differences, by shrinking the hypothesis space that a child is expected to entertain in the course of language maturation. Secondly, it supports the view that the speaker's internalised grammar is a dynamic system whose stable states result from the context-based interaction of several constraints. There is growing consensus on two major properties of these constraints: i) they are probabilistic "soft constraints" (Bresnan et al. 2001), and ii) they have an inherently functional nature, involving different types of linguistic (and non-linguistic) information (syntactic, semantic, etc.). 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. Last but not least, the probabilistic hypothesis prompts a methodological point. Differences in frequency distributions are not to be taken at face value, but must be subjected to probabilistic interpretations. Corpus evidence could in principle be the result of several interfering factors, be they grammatical or extra-grammatical. Often, the same evidence is compatible with different probabilistic models. In assessing different interpretations, we should thus be careful in assessing different probabilistic, and ultimately explanatory, models.

In the remainder of this paper we would like to show that this second hypothesis is not only more interesting and explanatorily adequate: it also finds independent confirmation in some experiments conducted with computational probabilistic language models that are trained automatically on annotated corpus data. In doing so, we shall exploit a powerful information theoretic principle known as Maximum Entropy (Ratnaparkhi, 1998). In the following section we first summarise the linguistic problem of subject/object understanding in Czech and Italian, to then move to an intuitive description of the Maximum Entropy (hereafter MaxEnt for short) framework.

2. Subjects and objects in Czech and Italian

Grammatical relations – such as subject (*S*) and direct object (*O*) – are 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 cues jointly co-operating in making a particular noun phrase 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 cue for *S* and *O* identification (henceforth *SOI*), 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. Moreover, there are languages where semantic and pragmatic constraints such as animacy and/or definiteness play a predominant role in the processing of grammatical relations. 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).

In the present paper we intend to probe this wide space of grammar variation through careful assessment of the distribution of *S* and *O* tokens in Italian and Czech, based on two syntactically annotated corpora: the *Prague Dependency Treebank* (PDT, Bohmová et al. 2003) for Czech, and the *Italian Syntactic Semantic Treebank* (ISST, Montemagni et al. 2003) for Italian. The 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 (they both adopt a dependency-based annotation scheme).

Czech and Italian provide an interesting vantage point for the cross-lingual analysis of grammatical variation. On the one hand, they appear to share two crucial features: 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 due to: the virtual non-existence of case marking in Italian (with the only marginal exception of personal pronouns), and the degree of phrase-order freedom in the two languages, already shown in Table 1 above. It is important to appreciate at this juncture that figures in Table 1 are indeed comparable from our perspective. Although it can be argued that the probability of object left-dislocations in Italian is underestimated because we only sampled journalistic prose, the difference from Czech distributions is indeed statistically significant because *both* corpora contain written texts only. We thus suggest that there is clear empirical evidence in favour of a systematic, higher phrase-order freedom in Czech, arguably related to the well-known correlation of Czech constituent placement with sentence information structure, 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, were not taken into account, as they happen not to be marked-up in the Italian corpus. Let us now move on to considering more subtle distributional patterns.

		Czech		Italian	
		Subj	Obj	Subj	Obj
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 2. *Distribution of Czech and Italian S and O wrt 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 3. *Distribution of Czech S and O wrt case*

According to the data reported in Table 2, Czech and Italian show similar correlation patterns 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. Italian shows a prominent 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 may play 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 are the strongest and most directly accessible cues for *SOI*, as they are marked both overtly and locally. This is also confirmed by psycholinguistic evidence, showing that subjects tend to rely on these cues to identify *S/O*. However, it should be observed that agreement can be relied upon conclusively 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'). Conversely, when N and V share the same person and number, no conclusion can be drawn, as trivially shown by a sentence like *il bambino legge il libro* 'the child reads the book'. In ISST, more than 58% of *O* tokens agree with their governing V, thus being formally indistinguishable from *S* on the basis of agreement features. PDT also exhibits a similar ratio, with 56% of *O* tokens agreeing with their verb head. Analogous considerations apply to case marking, whose perceptual reliability is undermined by morphological syncretism, whereby 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* tokens extracted from PDT are formally indistinguishable from *S* in case ending. Similarly, 45% of *S* tokens are formally indistinguishable from *O* uses on the same ground. All in all, this means that in 50% of the cases a Czech noun cannot be understood as the *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 Czech *SOI* is governed by a complex interplay of probabilistic constraints of a different nature (morpho-syntactic, semantic, word order, etc.) as the latter are neither singly necessary nor jointly sufficient to attack the processing task at hand. It is tempting to hypothesize that the joint distribution of these data can provide a statistically reliable basis upon which relevant probabilistic constraints are bootstrapped and combined consistently. This should be possible due to i) the different degrees of cue salience in the two languages and ii) the functional need to minimize processing ambiguity in ordinary communicative exchanges. With reference to the latter point, for example, we may surmise that a speaker will be more inclined to violate one constraint on *S/O* distribution (e.g. word order) when another cue is available (e.g. animacy) that strongly supports the intended interpretation only. The following section illustrates how a MaxEnt model can be used to model these intuitions by bootstrapping 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.

3. MaxEnt modelling

3.1. Linguistic representations as feature sets

Before we get into the mathematical nitty-gritty of MaxEnt modelling, it is useful to consider the problem of turning the linguistic structure of a piece of text into a feature-based representation that is amenable to a probabilistic treatment. A simple example of this transformation is shown in Table 4 below for the Italian sentence *Lo scampato pericolo scatena la squadra* ('the avoided danger rouses the team'). Words are arranged vertically in the second column of Table 4, whereas the first column specifies the corresponding numerical identifiers reflecting their order in the sentence. For each word, the corresponding row conveys a wide variety of linguistic information, ranging from its lemma, part of speech (POS) category and morpho-syntactic features such as person, number and gender (MS FEATS). The last two columns carry information about the dependency relations holding between the words in the sentence: the column "Head" contains the numerical identifier of the head of the current word token, whereas the column "Rel" specifies the relation type holding between the current word and its Head. For instance, the third word token in the sentence – *pericolo* – is the subject (sogg) of the verbal head *scatena* which is the fourth token in the sentence. *Pericolo* is in its turn modified by an adjective, *scampato*: this is encoded in the raw describing the token *scampato* where it is stated that it relates as a modifier (mod) to the nominal head identified by 3. The root of the sentence, corresponding to *scatena* in the case at hand, is marked by 0 in the "Head" column and by ROOT in the "Rel" column. All these pieces of linguistic information are technically represented through features.

Table 4 allows us to conceptualize the problem of assigning structure to a text as the task of assigning features to an ordered sequence of words, or word classification task. Now, stochastic models are particularly well suited to modelling classification tasks of this kind. In fact, this is at the basis of the idea of using a stochastic classifiers to tackle the problem of SOI. Turning back to Table 4, SOI is equivalent to assigning to the noun token *pericolo* a complex feature, consisting of a dependency relation ("sogg" in the case at hand) and a numerical identifier, pointing to the verbal head (*scatena*) with respect to which the subject relation obtains.

ID	Form	Lemma	POS	MS FEATS	Head	Rel
1	Lo	Lo	RD	gen=M|num=S	3	det
2	scampato	scampato	A	gen=M|num=S	3	mod
3	pericolo	pericolo	S	gen=M|num=S	4	sogg
4	scatena	scatenare	V	num=S|per=3|mod=I|tmp=P	0	ROOT
5	la	la	RD	gen=F|num=S	6	det
6	squadra	squadra	S	gen=F|num=S	4	ogg_d
7	.	.	PU	_	6	punc

Table 4. *A feature-based representation of a sentence structure.*

3.2. A probabilistic interpretation of SOI

The *SOI* problem is now easily amenable to a probabilistic formulation. We have to gauge how probable it is for the word *pericolo* to be the subject (or object) of *scatena*, on the basis of some given linguistic information about *pericolo* and its context. The information may include inherent features of the word *pericolo* itself, namely that it is a masculine singular noun and has a certain meaning, and its relation to the embedding context: *pericolo* precedes the verb, it agrees with it, it is used with a definite article etc.

More formally, we call C_{pericolo} a feature-based representation of *pericolo* plus its surrounding context. Looking back at Table 4, C_{pericolo} may correspond to a subset of the cells shaded in grey, including the definite article *il* and the verbal head. The probability of *pericolo* being the subject of our sentence can then be expressed as the conditional probability $p(pericolo_{\text{subj}}|C_{\text{pericolo}})$, which reads "the probability of *pericolo* being the subject, given a specific featural representation of its embedding context". Since both $pericolo_{\text{subj}}$ and C_{pericolo} are (set of) features, we call $pericolo_{\text{subj}}$ the target feature, and C_{pericolo} the set of linguistic cues on whose basis the target feature is assigned. In $p(pericolo_{\text{subj}}|C_{\text{pericolo}})$, C_{pericolo} is used as a probabilistic constraint conditioning the probability of *pericolo* being the subject. But how can we estimate $p(\text{subj}|C_{\text{pericolo}})$?

Before we answer this question, let us shortly consider first how we tackle SOI as a probabilistic classification task. This corresponds to finding out the most probable subject candidate in the context given. If $p(pericolo_{\text{subj}}|C_{\text{pericolo}})$ is higher than $p(pericolo_{\text{obj}}|C_{\text{pericolo}})$ then *pericolo* is interpreted as a subject. The object interpretation prevails in the opposite case. To be more concrete, suppose that the probability that *pericolo* be the object of a non-agreeing main verb is smaller that the probability of being a subject of the same verb: $p(pericolo_{\text{obj}}|\text{non-agr}) > p(pericolo_{\text{subj}}|\text{non-agr})$. On the basis of our criterion, *pericolo* will always be given an object interpretation in non-agreeing contexts.

There is a potentially serious problem here, however. If $p(pericolo_{\text{subj}}|\text{non-agr})$ has a non zero probability to occur, this means that we can possibly find contexts where a non-agreeing noun has a subject interpretation. How does our probabilistic classifier deal with these somewhat exceptional cases? Will it always get them wrong? Like rule-based systems, probabilistic classifiers can deal with marginal or exceptional cases as long as we are able to identify what relevant features are responsible for them. If our feature-based representation includes some other information than (non)-agr, and if this information has a bearing on the cases we want to capture, then we can make our probabilistic model sensitive to it and deal with the intended exceptions. Otherwise, the more general interpretation will always prevail. This excursus emphasizes an important aspect of probabilistic models in general, and MaxEnt models in particular: a fundamental prerequisite to building a linguistically meaningful model is to select those features that are relevant to the classification task. No probabilistic model can crank out sensible results on the basis of a misleading or incomplete representation of linguistic contexts.

3.3. MaxEnt estimation

One way to approximate $p(\text{subj}|C_{\text{pericolo}})$ is by counting, in a suitably annotated corpus, how many times a subject is found in contexts matching the C_{pericolo} featural representation, divided by how many times we find C_{pericolo} matching contexts overall. To be more concrete, let us consider the simple (made up) example in Table 5.a) below, showing the distribution of 14 subjects and 24 objects according to a four-way classification of their embedding contexts: presence vs absence of agreement and pre- vs post-verbal position. As there are 38 attested contexts overall, matching our feature combination, one can say that the probability of finding a subject in contexts where a noun concords with an ensuing main verb is 5/19. Table 5.b) gives the estimated probabilities for all contexts.

a)

	agr		no-agr		
	pre-v	post-v	post-v	pre-v	
subj	10	3	0	1	14
obj	5	15	3	1	24
	15	18	3	2	38

b)

	Agr		no-agr	
	pre-v	post-v	post-v	pre-v
subj	5/19	3/38	0	1/38
obj	5/38	15/38	3/38	1/38

Table 5. *A made up example of distribution of subjects and objects classified wrt agreement and pre/post-verbal position.*

Table 5.b) is based on the reasonable assumption that the actual distribution in a reference corpus should reflect general trends in the interaction of linguistic factors. In fact, at a closer look, our calculations are based on the rather more restrictive (a priori) assumption that attested distributions are *directly* interpretable as a (probabilistic) model of interacting linguistic factors. For example, if we do not find evidence for a non-agreeing subject in post-verbal position, in Table 5.b) we come to the conclusion that this case has a null probability to occur. In fact, this conclusion may be too hasty. May be, our reference corpus was too small or biased for the linguistic factors prompting a non-agreeing subject to occur. After all, corpus information is inherently *incomplete*, as lexical data are known to be exceedingly sparse, with new events continuously popping up in an endless Zipfian tail. This brings us to the following questions: how should we interpret corpus distributions in a probabilistic framework? What kind of inferences are we allowed to make in the face of incomplete corpus evidence? The MaxEnt framework addresses all these questions in a maximally cautious way. The general philosophy is that we are not entitled to jump to premature conclusions. But what does this exactly mean?

In probabilistic terms, the most cautious a priori assumption that can possibly be entertained is that, all things being equal, data are distributed evenly. We do not expect things to occur in skewed ways. The faces of a dice are naturally assumed to have an equal probability to show up. Likewise, in a typological perspective, there is no a priori reason to assume that SVO languages should be more likely to occur than OVS languages are. Surely, things are not always equal. A dice can be loaded. The human language processing system may function is such a way that SVO patterns are easier to be understood and acquired. Be that as it may, the MaxEnt framework suggests that probabilistic models should not be more biased than required by input evidence. Their distance from the zero assumption (even distribution) should be made as small as necessary for the model to predict all attested data. By maximizing the entropy of a probabilistic distribution we exactly achieve this result: we minimise its distance from the even (or equiprobable) distribution. Turning back to Table 5.b), there is no need to say that the probability of a non-agreeing post-verbal subject is null in Italian: its probability must be such that, in all attested contexts, it is smaller that the probability of a non-agreeing post-verbal object. In a MaxEnt framework, making probabilities more skewed than necessary is avoided.

3.4. MaxEnt at work

The MaxEnt framework offers a mathematically sound way to build a probabilistic model for *SOI* which combines different linguistic cues. As we just saw, given a linguistic context *c* and an outcome $a \in A$ that depends on *c*, in the MaxEnt 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) $$p(a \mid c) = \frac{1}{Z(c)} \prod_{j=1}^{k} \alpha_j^{f_j(a,c)}$$

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 $\alpha_1, \ldots, \alpha_k$ 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 building the conditional probability distribution $p(\varphi|\sigma)$ of having a syntactic function φ in a syntactic context σ. Adopting the MaxEnt 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; a similar situation can be observed in PDT where complete subject-verb-object configurations occur in only 20% of the cases. Due to the comparative sparseness of canonical *SVO* constructions in Czech and 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.

3.5. 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 feature 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 allows 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 tests the position of the noun wrt the verb, for instance:

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

Animacy. This is the main semantic feature, which tests whether the noun in σ is animate or inanimate (cf. section 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 scales being 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 common nouns.

4. Testing the model

The Italian MaxEnt model was trained on 14,643 verb-subject/object pairs extracted from ISST. For Czech, we used a training corpus of 37,947 verb-subject/object pairs extracted from PDT. In both cases, the training set was obtained by extracting all verb-subject and verb-object dependencies headed by a verb used in the active voice, with the exclusion of all cases where the position of the nominal constituent was grammatically determined (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.

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. Evaluation was carried out 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. We have assumed a baseline score of 56% for Italian and of 66% for Czech, corresponding to the result yielded by a naive model assigning to each test pair the most frequent relation in the training corpus, i.e. subject. Experiments were carried out with the general features illustrated in section 3.5: verb agreement, case (for Czech only), word order, noun animacy and noun definiteness.

Accuracy on the test corpus is 88.4% for Italian and 85.4% for Czech. A detailed error analysis for the two languages is reported in Table 6, showing that in both languages subject identification appears to be particularly problematic. In Czech, it appears that the prototypically mistaken subjects are post-verbal (71.14%), inanimate (72.64%), ambiguously case-marked (70.65%) and agreeing with the verb (70.15%), where reported percentages refer to the whole error set. Likewise, Italian mistaken subjects can be described thus: they typically occur in post-verbal position (71.55%), are mostly inanimate (64.66%) and agree with the verb (61.21%). Interestingly, in both languages, the highest number of errors occurs when a) N has the least prototypical syntactic and semantic properties for *O* or *S* (relative to word order and noun animacy) and b) morpho-syntactic features such as agreement and case are neutralised. This shows that MaxEnt is able to home in on the core linguistic properties that govern the distribution of *S* and *O* in Italian and Czech, while remaining uncertain in the face of somewhat peripheral and occasional cases.

	Czech		**Italian**	
	Subj	Obj	Subj	Obj
Preverb	1.99%	19.40%	0.00%	6.90%
Postverb	71.14%	7.46%	71.55%	21.55%
Anim	0.50%	3.98%	6.90%	21.55%
Inanim	72.64%	22.89%	64.66%	6.90%
Nomin	0.00%	1.00%	Na	
Genitive	0.50%	0.00%		
Dative	1.99%	0.00%		
Accus	0.00%	0.00%		
Instrum	0.00%	0.00%		
Ambig	70.65%	25.87%		
Agr	70.15%	25.87%	61.21%	12.07%
NoAgr	2.99%	0.50%	7.76%	1.72%
NAAgr	0.00%	0.50%	2.59%	14.66%

Table 6. *Types of errors 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 7. *Feature value weights for Czech and Italian.*

A further way to evaluate the goodness of fit of our model is by inspecting the weights associated with feature values for the two languages. They are reported in Table 7, 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 to the subject relation. For Czech, nominative case is strongly associated with subjects while the other cases with objects. Moreover, in both languages preverbal subjects are strongly preferred over preverbal objects; animate subjects are preferred over animate objects; pronouns and proper names are typically subjects.

Let us now try to relate these feature values to the Markedness Hierarchies reported in section 3.5. Interestingly enough, if we rank the Italian *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 section 3.5. 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*.

4.1. Evaluating comparative 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 MaxEnt can successfully be applied to learn constraint rankings in Optimality Theory, by assuming the parameter weights <α1, …, αk> as the ranking values of the constraints.

Table 8 illustrates constraint rankings for the two languages, ordered by decreasing weight values for both *S* and *O*. Note that, although not all constraints are applicable in both languages, the weights associated with equally applicable constraints exhibit the same relative salience in Czech and Italian. This seems to suggest existence of a rather dominant (if not universal) salience scale of *S* and *O* processing constraints, in spite of the considerable difference in the marking strategies adopted by the two languages. As the relative weight of each constraint crucially depends on its overall interaction with other constraints on a processing task, absolute weight values can considerably vary from language to language, with a resulting impact on the distribution of *S* and *O* constructions. For example, the possibility of overtly and unambiguously marking a direct object with case inflection makes wider room for preverbal use of objects in Czech. Conversely, lack of case marking in

Italian considerably limits the preverbal distribution of direct objects. More importantly, this evidence appears to be an epiphenomenon of the interaction of fairly stable and invariant preferences, reflecting common functional tendencies in language processing. As shown in Table 8, if constraint ranking largely confirms the interplay between animacy and word order in Italian, Czech does not contradict it but rather re-modulate it somewhat, due to "perturbation" factors introduced by its richer battery of case markers. The range of phenomena presented here sheds light on the complex interaction of semantically interpretable and uninterpretable features in different languages, and on the impact that comparatively small differences in the range of features may have on the overall grammatical patterns exhibited by those languages.

Constraints for S			Constraints for O		
Feature	Italian	Czech	Feature	Italian	Czech
Preverbal	1.31E+00	1.24E+00	*Genitive*	Na	1.51E+00
Nomin	Na	1.23E+00	***NoAgr***	1.58E+00	1.50E+00
Agr	1.28E+00	1.18E+00	*Dative*	Na	1.49E+00
Anim	1.28E+00	1.16E+00	*Accus*	Na	1.39E+00
Inanim	8.16E-01	1.03E+00	*Instrum*	Na	1.39E+00
Postverbal	5.39E-01	8.77E-01	***Postverbal***	1.38E+00	1.17E+00
Genitive	Na	2.94E-01	***Inanim***	1.23E+00	9.63E-01
NoAgr	1.52E-01	7.71E-02	***Agr***	4.67E-01	6.67E-01
Dative	Na	2.85E-02	***Anim***	3.17E-01	6.63E-01
Accus	Na	8.06E-03	***Preverbal***	2.11E-02	5.40E-01
Instrum	Na	3.80E-03	*Nomin*	Na	2.22E-02

Table 8. *Ranked constraints for S and O identification in Czech and Italian.*

5. Concluding remarks

Probabilistic language models, machine language learning algorithms and linguistic theorizing all appear to support a view of language processing as a the result of dynamic, on-line resolution of conflicting grammatical constraints. We begin to gain considerable insights into the nature and behaviour of these constraints upon observing their actual distribution in perceptually salient contexts. The approach allows scholars to investigate 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).

In this paper we hope to have convincingly showed that the current availability of comparable, richly annotated corpora and of mathematical tools and probabilistic models for corpus exploration make the 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: first, it is bound to shed light on the integration of performance and competence factors in language study. Secondly, it will make mathematical models of language increasingly able to accommodate richer and richer language structures, thus putting explanatory theoretical accounts to the test of a usage-based empirical verification.

In particular, we put to empirical test the hypothesis that large surface differences in the grammatical patterns of Czech and Italian can in fact be brought down to small differences in the absolute strength values of competing grammatical constraints. More importantly, value differences are such that the relative ranking of constraints that hold in the two languages is in fact invariant. This lends support to a universalistic interpretation of ranking. Both Czech and Italian subjects comply with anymacy and definitess markedness hierarchies and both show a similar tendency to occur before the verb. Nonetheless, their distribution in the two languages is dramatically different due to the interference of a semantically uninterpretable feature like case, which plays a prominent role in Czech, but is marginal if not irrelevant in Italian.

Our conclusion flies in the face of the common belief that probabilistic models are only surface accounts of data distributions. In fact, non trivial models of this kind can be used as a probe into the space of grammar competence. Moreover, computer simulations of the dynamics of grammar constraints in language processing are by no means incompatible with the view that speakers internalize a complex body of abstract linguistic competence. We contend that such a body of abstract knowledge is more intimately related to usage-based aspects of the language input than some linguists have so far been ready to recognize.

References

Aissen, J. 2003. Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory* 21: 435-483.

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

Berger A., S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22 (1): 39-71.

Bohmová, A., J. Hajič, E. Hajičová, and B. Hladka. 2003. The Prague Dependency Treebank: Three-Level Annotation Scenario. In: A. Abeille (ed.), *Treebanks: Building and using syntactically annotated corpora*, 103-128. Dordrecht: Kluwer.

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

Burzio, L. 2005. Sources of paradigm uniformity. In: L. J. Downing, T. A. Hall, and R. Raffelsiefen (eds.), *Paradigms in Phonological Theory*, 65-106. Oxford: OUP.

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*, 107-132. Amsterdam: Benjamins.

Croft, W. 2003. *Typology and Universals*. Second edition, Cambridge: CUP.

Goldwater, S., M. Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In: J. Spenader, A. Eriksson, and Ö. Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111-120. April 26-27, 2003, Stockholm University.

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: J. Dekkers, F. van der Leeuw, and J. van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*, 88-120. Oxford: OUP.

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, and A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: OUP.

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

Matthews, D., E. V. Lieven, A. L. Theakston, M. Tomasello. 2005. The role of frequency in the acquisition of English word order. *Cognitive Development* 20: 121-136.

MacWhinney B., E. Bates, and R. Kliegl. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior* 23: 127-150.

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

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

Pala, K., P. Smrz. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology* 7 (1/2): 79-88.

Ratnaparkhi, A. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. Dissertation, University of Pennsylvania.

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

Stone, G. 1990. Czech and Slovak. In: B. Comrie (ed.), *The World's Major Languages*. New York: OUP.