Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order

Chiara Alzetta[+], Felice Dell'Orletta[*], Simonetta Montemagni[*], Giulia Venturi[*]
[+] Università degli Studi di Genova, Via Balbi 5, Genova, Italy
[*]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) – ItaliaNLP Lab, *www.italianlp.it*
via G. Moruzzi, 1- Pisa, Italy
chiara.alzetta@edu.unige.it, {felice.dellorletta,simonetta.montemagni,giulia.venturi}@ilc.cnr.it

Abstract

The paper presents a new methodology aimed at acquiring typological evidence from "gold" treebanks for different languages. In particular, it investigates whether and to what extent algorithms developed for assessing the plausibility of automatically produced syntactic annotations could contribute to shed light on key issues of the linguistic typological literature. It reports the first and promising results of a case study focusing on word order patterns carried out on three different languages (English, Italian and Spanish).

Keywords: Linguistic Knowledge Extraction, Dependency Treebanks, Linguistic Typology

1. Introduction

The interaction between linguistics and computational linguistics has a long history dating back to the 60's. In Kučera (1982), it is explicitly stated that "computational linguistics provides important potential tools for the testing of theoretical linguistic constructs and of their power to predict actual language use". This still appears to represent a key objective, as claimed e.g. by Martin Kay in his ACL Lifetime Award speech in 2005 (Kay, 2005), or by the more recent papers gathered in the Special Issue of the journal "Linguistic Issues in Language Technology" (LiLT) focusing on the relationship between language technology and linguistic insights (Baldwin and Kordoni, 2011). After more than 40 years from the first pioneering studies, the growing availability of linguistic resources such as annotated corpora for many languages combined with the increasing reliability of Natural Language Processing (NLP) methods and tools enables the acquisition of quantitative evidence ranging across different levels of linguistic description which can significantly contribute to the study of open issues of the theoretical linguistic literature.

This holds particularly true for the area of typological studies which can benefit a lot from this synergy, making it possible to acquire quantitative evidence shedding light on how, why and to what extent languages vary with respect to key features covering major areas of language structure. By exploring collections of linguistically annotated corpora for different languages, complex and articulated frequency distributions of language constructions can be extracted. Information acquired from available corpora can significantly enrich typological descriptions of languages such as the *World Atlas of Language Structures* (WALS) (M. S. Dryer and M. Haspelmath, 2013) [1], the most commonly-used and broadest database of structural (phonological, grammatical, lexical) properties of languages whose data are based on primary sources such as grammars, dictionaries and scientific papers. But impact and role of this information type cannot be limited to the descriptive level. Typological evidence inferred from linguistically annotated corpora for different languages can significantly contribute to model linguistic variation within and across languages. Word order variation represents a widely investigated topic of the typological literature whose recent developments include fine-grained studies based on a wide range of features and their frequency distributions typically acquired from annotated corpora (O'Horan et al., 2016). To mention only a few, see e.g. Gulordava and Merlo (2015a), Gulordava and Merlo (2015b) Futrell et al. (2015), Merlo (2016).

More recently, such an approach to typological studies has also been prompted by the availability of multi–lingual treebanks such as those designed and constructed within the Universal Dependencies project[2] for over 50 languages. Universal Dependencies (UD) is a framework for cross–linguistically consistent treebank annotation aiming to capture similarities as well as idiosyncrasies among typologically different languages (e.g., morphologically rich languages, pro–drop languages, and languages featuring clitic doubling). The goal in developing UD was not only to support comparative evaluation and cross–lingual parsing but also to enable comparative linguistic studies (Nivre, 2015). Within this area of research, the paper reports the results of preliminary experiments aimed at acquiring quantitative typological evidence from a selection of the UD treebanks for different languages. In particular, it focuses on the widely investigated topic of word order, with the specific aim of reconstructing word order patterns within and across languages, for what concerns both the linear order of words and its degree of flexibility (giving rise to a wide typology of languages going from fixed-order to free word order languages). For the acquired word order patterns, the study also aims at investigating the factors underlying the preference for one or the other order, both intra- and cross-linguistically.

To pursue this goal, we decided to test whether existing algorithms for assessing the plausibility of automatically produced syntactic annotations could be used to acquire useful quantitative typological evidence. In fact, the result of these algorithms is typically driven by linguistic properties

[1] Available online at http://wals.info

[2] http://universaldependencies.org/

characterizing the language being processed: by comparing the results achieved against different languages, it is possible to acquire information concerning typological similarities and differences. This kind of algorithms operate at the level of either the whole syntactic tree (cfr. for example Dell'Orletta et al. (2011) and Reichart and Rappoport (2009)), or individual dependencies (see, among others, Dell'Orletta et al. (2013) and Che et al. (2014)). Given the focus of this study on specific constructions, we selected the class of algorithms operating at the level of individual dependencies, and in particular on those ranking dependencies by decreasing plausibility of annotation. These algorithms, originally meant to discern reliable from unreliable annotations within the automatic output of parsers, have also been applied to manually revised (i.e. "gold") linguistic annotations with the final aim of identifying annotation inconsistencies, and thus for also detecting annotation errors (Alzetta et al., 2018). Tusa et al. (2016) represent the first attempt to exploit the plausibility score returned by this class of algorithms to acquire linguistic evidence, i.e. to infer the prototypicality degree of specific linguistic constructions. The experiment was carried out against the *Italian Universal Dependency Treebank* (IUDT) (Bosco et al., 2013) with promising results: the plausibility-based ranking of dependencies corresponding to specific syntactic constructions turned out to closely reflect their linguistic "markedness" degree.

In what follows, we focus on word pattern variation across three different languages, English, Italian and Spanish. This goal is pursued by applying a plausibility assessment algorithm against the UD treebanks available for these languages. Achieved results have been compared with the threefold aim of: i) reconstructing the frequency distributions of different word order patterns, with particular attention to specific constructions (Subject-Verb and Adjective-Noun); ii) assessing similarities and differences across languages; and iii) identifying and weighting the factors underlying the different word order patterns identified.

2. Background and Motivation

Starting from Greenberg (1963), word order has been used to set up a typology of languages based on the notion that "certain languages tend consistently to put modifying or limiting elements before those modified or limited, while others just as consistently do the opposite". Within this area of research, the relative ordering of constituents at the clausal level (e.g. verb and subject) as well as at the phrasal level (e.g. noun and modifying adjective) has been widely investigated in the typological literature. Nowadays, the outcome of these studies has been collected in publicly-accessible typological databases. Table 1 reports - for the three languages considered in our study - the different orderings of lexical (i.e. non pronominal) Subject and Verb, and of Adjective and Noun as resulting from the *World Atlas of Language Structures* (or WALS) and the *Syntactic Structures of the World's Languages* (or SSWL) (SSWL, 2009) databases. It can be noted that the two provide a slightly different picture for the three languages, maybe following from the fact that whereas the former records the "dominant" order the latter testifies "productive" word order patterns, which may not always coincide. According to WALS, no dominant Subject-Verb order exists for both Italian and Spanish (M. S. Dryer, 2013b), whereas Subject-Verb is considered a productive word order for both languages in SSWL. If Adjective-Noun is a productive order for all the three languages according to SSWL, in WALS the Noun-Adjective order is considered dominant in Italian and Spanish while the reverse order is reported for English (M. S. Dryer, 2013a).

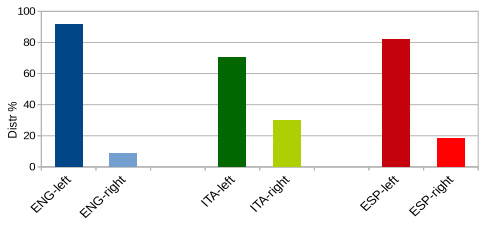

Figure 1: Distribution of right– vs left–headed non–pronominal *nsubj* relations in the three UD treebanks.

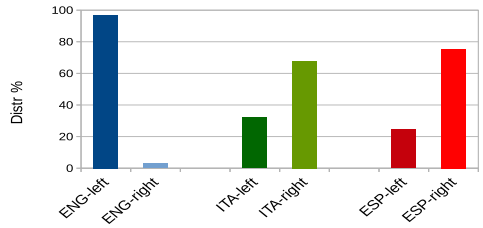

Figure 2: Distribution of right– vs left–headed *amod* relations in the three UD treebanks.

Consider now the picture emerging from actual usage as attested in the English, Italian and Spanish UD treebanks. Figures 1 and 2 report, for the three treebanks, the percentage distribution of different word orders involving non pronominal Subjects and Verbs (corresponding to the *nsubj* dependency relation), and Adjectives and Nouns (*amod*), respectively. Note that *Left* and *Right* in the figures refer to the position of the dependent (subject or adjectival modifier) with respect to its syntactic head. As it can be noted, all languages turned out to prefer the Subject-Verb order, but with significant differences: namely, the Italian and Spanish treebanks are characterized by a much higher percentage of left–headed subjects than English (i.e. 30% in Italian, 18% in Spanish and 8% in English). For what concerns adjectival modifiers, Figure 2 shows that whereas for English the Adjective-Noun order is highly preferred (though not the only possible one) in Italian and Spanish the reverse order is rather preferred, i.e. with the adjective occurring on the right side of the head.

Corpus-based evidence helps quantifying attested word order patterns across languages, thus leading to a more articulated picture of word order variation. Registered order variants, however, do not appear to be evenly distributed, both intra- and cross-linguistically. Some are used more often and show less grammatical or stylistic restrictions than others. Let us consider, for example, the relative ordering of

WALS				
	Word Order	**English**	**Italian**	**Spanish**
Subject-Verb	Subject-Verb	+	No dominant order	No dominant order
	Verb-Subject	-	No dominant order	No dominant order
Adjective-Noun	Adjective-Noun	+	-	-
	Noun-Adjective	-	+	+
SSWL				
Subject-Verb	Subject-Verb	+	+	+
	Verb-Subject	-	+	+
Adjective-Noun	Adjective-Noun	+	+	+
	Noun-Adjective	+	+	+

Table 1: Subject-Verb and Adjective-Noun order in WALS and SSWL.

Subject and Verb. Figure 3 reports different *nsubj* instances occurring in the Italian UD treebank[3]. The subject in a) (lit. 'In this case, *he* answers within limits ...') and c) (lit. '... *the rights not overtly granted by the licensor* remain confidential') occurs in the same position, i.e. pre-verbally. There are however important differences worth noting here: in a) a pronominal subject immediately precedes the verb, while in c) a long–distance dependency relation links the nominal subject to its head. The question which naturally arises here is how a) and c) relate to each other, and what are the underlying properties explaining this difference, if any. On the other hand, the sentence reported in b) (lit. 'Here, from each corner of the world, arrive *300 thousand patients*') exemplifies a different, less common, Verb-Subject order involving a nominal subject. The question at this point is how a) and c), both with pre-verbal subjects, relate to b), with a post-verbal subject. Both questions cannot be answered by simply considering finer-grained frequency distributions of different types of ordering. On the basis of this, we can claim that a) and c) represent more likely verb-subject instances than b). But this may not be the case. To answer questions like these, the properties underlying the different word order patterns in a given language and cross-linguistically need to be investigated.

Thanks to the availability of corpora for different languages with manually revised linguistic annotation, the focus of studies on word order variation across languages moved from discerning possible vs impossible word orders as in the pioneering studies, to defining dominant vs rare word order patterns based on actual frequencies attested in corpora. More recently, thanks to the wide variety of features which can be tracked down and quantified in linguistically annotated corpora, current explanations of word order variation can also aim at capturing finer-grained distinctions able to predict the frequency distribution of attested word orders in different languages. In what follows, we will try to exploit the wide range of features which can be extracted from treebanks not only to characterize word order patterns across languages, but also to identify and weight the factors underlying them.

3. Method and Data

The methodology we devised to acquire typological evidence from gold treebanks is based on the parse plausibility assessment algorithm named LISCA (LInguiStically–driven Selection of Correct Arcs) (Dell'Orletta et al., 2013). As illustrated in details in Section 3.1., the algorithm exploits statistics about a wide range of linguistic features (covering different description levels, going from raw text to morpho-syntax and dependency syntax) extracted from a large reference corpus of automatically parsed sentences and uses them to assign a *plausibility* score to each dependency arc contained in a target corpus belonging to the same variety of use (i.e. textual genre) of the automatically parsed corpus. Accordingly, all the arcs contained in the target corpus are ranked from those characterized by a high LISCA score to arcs with lower scores: the higher the score, the more similar the linguistic context of an arc with respect to the statistics acquired from the large reference corpus. The underlying assumption is that syntactic structures that are more frequently generated by a parser are more likely to be plausible than less frequently generated ones.

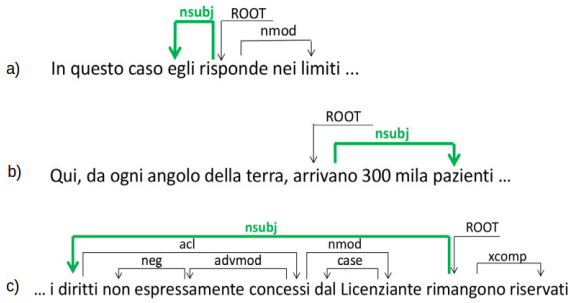

Figure 3: Different instances of the *nsubj* relation in the Italian UD treebank.

3.1. The LISCA Algorithm

LISCA takes as input a set of parsed sentences and it assigns a plausibility score to each dependency, which is defined as a triple (d, h, t) where d is the dependent, h is the head, and t is the type of dependency connecting d to h. The algorithm operates in two steps: *1)* it collects statistics about a set of linguistically motivated features extracted from a dependency annotated corpus obtained through automatic dependency parsing, and *2)* it combines the feature statistics extracted from the corpus used during the previous step. The final plausibility score associated with a given dependency arc results from the combination of the weights associated with these features: the score is computed as a

[3] In the examples, the dependent is italicized and the head underlined.

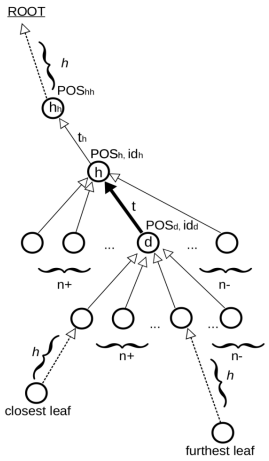

Figure 4: Features used by LISCA to measure arc(d, h, t) plausibility.

simple product of the individual feature weights.

Figure 4 illustrates the features taken into account by LISCA for measuring the plausibility of a given syntactic dependency (d, h, t). For the purposes of the present study, LISCA has been used in its de–lexicalized version in order to abstract away from variation resulting from lexical effects. In particular, two different types of features are considered:

- *local* features, corresponding to the characteristics of the syntactic arc considered, such as the distance in terms of tokens between d and h, or the associative strength linking the grammatical categories (i.e. POS$_d$ and POS$_h$) involved in the relation, or the POS of the head governor and the type of syntactic dependency connecting it to h;
- *global* features, aimed at locating the arc considered within the overall syntactic structure of the sentence: for example, the distance of d from the root of the tree, or from the closest or most distant leaf node, or the number of "brothers" and "children" nodes of d, occurring respectively to its right or left in the linear order of the sentence.

LISCA was successfully used against both the output of dependency parsers and gold treebanks. While in the first case the plausibility score was meant to identify unreliable automatically produced dependency relations, in the second case it was used to detect shades of syntactic markedness of syntactic constructions in manually annotated corpora. The latter is the case of Tusa et al. (2016), where the LISCA ranking was used to investigate the linguistic notion of "markedness" (Haspelmath, 2016): a given linguistic construction is considered "marked" when it deviates from the "linguistic norm", i.e. it is "abnormal". Accordingly, unmarked constructions are expected to be characterized by higher LISCA scores and – conversely – constructions characterized by increasing degrees of markedness are associated with lower scores.

Let us go back to the different instances of the *nsubj* relation in Figure 3. The plausibility score assigned to them by LISCA results in the following ranking: a) is assigned a higher score with respect to b), whose score in turn is higher than that assigned to c). This ranking follows from the combination of both local and global features taking into account the overall tree structure. From the resulting LISCA score, it turned out e.g. that in Italian longer distance subjects are less prototypical than post–verbal and shorter ones.

4. Languages and Corpora

For the specific concerns of this study we focused on two typologically close languages, namely Italian and Spanish, and a more distant one, English: for what concerns morphology, all three languages are fusional, although English has very few inflectional morphemes, which makes it rather similar to isolating languages. These properties imply that the prototypical sequence of the main constituents in English strictly follows the linear order Subject-Verb-Object (SVO); Italian and Spanish are SVO languages too, but show more syntactic freedom in the linear ordering of constituents. Because of these properties, highly related to the linguistic type each language belongs to, we expect to observe a similar behaviour for typologically close languages and, on the other hand, significant differences in case of typologically distant languages.

The corpora used to collect the statistics to build the LISCA models (step 1 in Section 3.1. above) are represented by the English, Italian and Spanish Wikipedia, for a total of around 40 million tokens for each language. The Spanish and English corpora were morpho–syntactically annotated and parsed by the UDPipe pipeline (Straka et al., 2016) trained on the Universal Dependency treebanks, version 2.0 (J. Nivre and A. Željko and A.Lars and et alii, 2017). The Italian corpus was morpho–syntactically tagged using the ILC–POS–Tagger (Dell'Orletta, 2009), and then parsed with UDPipe.

LISCA, trained on the models we created earlier, was then applied to the Italian, English and Spanish UD Treebanks in order to assign a plausibility score to each dependency relation. The English Web Treebank (Silveira et al., 2014) contains 16,624 sentences and 254,830 tokens, while the Italian Universal Dependency Treebank (Bosco et al., 2013) contains 13,815 sentences corresponding to 325,816 tokens. The Spanish UD treebank (McDonald et al., 2013) is the smallest one, with 4,000 sentences and 112,718 tokens. For all resources, most part of the sentences comes from blogs and/or newspapers.

5. Data Analysis

For each treebank, the dependencies were first ordered by decreasing LISCA scores. The list of ordered dependencies was then subdivided into 10 groups, henceforth "bins", each corresponding to 10% of the total. The distribution of syntactic dependencies in the LISCA bins was analyzed in order to investigate whether and to what extent it could be used to acquire typological trends, i.e. similarities and differences across languages. This analysis has been carried out by comparing the dependency rankings by LISCA (each subdivided into 10 bins) for the three languages taken into account. As reported by O'Horan et al. (2016), the

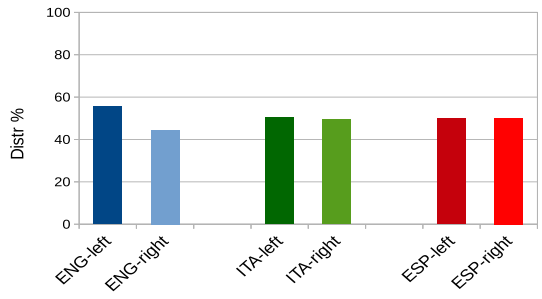

Figure 5: Frequency distribution of right– vs left–headed dependencies in the three UD treebanks.

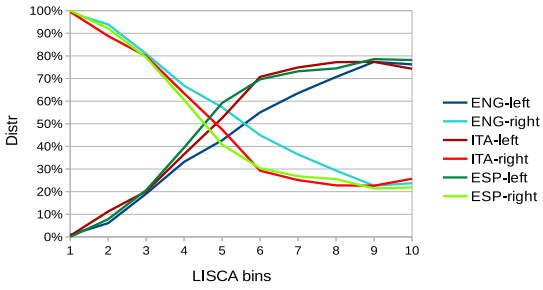

Figure 6: Distribution of right– vs left–headed dependencies across the bins.

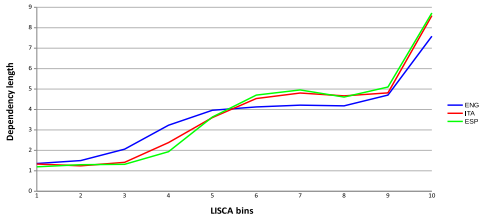

Figure 7: Average length of dependencies across the bins.

automatic learning of typological information is typically carried out from parallel texts. In the case of our study, parallelism is not concerned with texts, but rather with the ranking of instances of dependendency relations by LISCA: besides the fact that the LISCA score is based on the same set of properties for the three languages, comparability is guaranteed here by the same inventory of dependency relations and annotation guidelines shared by the UD treebanks taken into account.

In what follows, we will focus on word order patterns: Section 5.1. focuses on a cross-lingual analysis of general trends of word order formalized in terms of the direction of the dependency link, whereas Section 5.2. reports the results of an in–depth analysis of the frequency distribution of two dependency relations, i.e. nominal subject (*nsubj*) and adjectival modifier (*amod*), corresponding to widely investigated constructions in the typological literature.

5.1. Word Order Patterns: General Trends

As a first step, for each treebank we analyzed how the dependencies are distributed with respect to their direction. For the specific purposes of this study, the order is defined by the right or left direction of the dependency that connects d to h with respect to the linear order of words in the sentence. Figure 5 shows the frequency distribution of all dependencies in the three considered UD treebanks, distinguishing right–headed dependencies (i.e. $d > h$) vs left–headed ones (i.e. $h < d$).[4] Interestingly enough, in the three treebanks the frequency of the the two word orders is similar: whereas for Italian and Spanish they are equally partitioned (50% both $d > h$ and $h < d$), for English the ordering $d > h$ covers 55% of the cases. Yet, if we focus on the distribution of dependencies across the LISCA bins interesting differences across languages can be observed (see Figure 6). Despite their similar frequency in the three languages, right– vs left–headed dependencies are described by opposite trends: whereas in the first bins $d > h$ dependencies are more frequent, in the latter $h < d$ dependencies are mainly observed.

This result confirms the intuition we started from, i.e. that statistics about the frequency distribution of right- vs left-headed dependencies in treebanks do not say much about the underlying structural properties of languages. On the contrary, the direction-based distribution of dependencies resulting from the LISCA ranking shows interesting similarities and differences across languages. All languages share the same trend, i.e. the top LISCA (namely 1-3) bins mainly contain right-headed dependencies, whose occurrence progressively decreases across the bins, until the last two bins (9-10) where they represent around the 20-25% of the dependencies. Similar observations hold in the case of left-headed dependencies, which are characterized by the reverse trend: their occurrence starts in the second bin and progressively increases to cover about 80% of the relations in the last two bins (9-10). Although this trend is shared by the three languages taken into account, there are also significant differences worth being pointed out. In fact, the trend reported in Figure 6 for English is slightly different if compared with what observed for the other two Romance languages. For English, the lines representing the distribution of the left– and right–headed dependencies are less steep and cross at the level of the 6th bin, while the Italian and Spanish lines are steeper and cross in the 5th one. This is to say that the distribution of dependencies in the interval between the 4th and 9th bins is language-specific. A possible explanation for this is that English word order is more rigid and shows a higher number of right–headed dependencies (including explicit subjects). On the contrary, Italian and Spanish, characterized by a more flexible word order, show a higher variability at the level of dependency direction.

We can now try to understand what distinguishes right– or left–headed dependencies occurring in the top vs bottom bins. Consider the distribution of dependencies with respect to their length, as shown in Figure 7: all languages share the fact that longer dependencies are ranked

[4] In all figures *left* and *right* refer to the position of d relative to its governor h, respectively corresponding to right–headed (i.e. $d > h$) vs left–headed (i.e. $h < d$) dependencies.

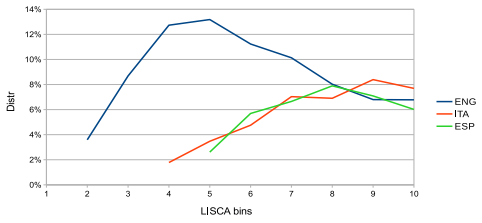

Figure 8: Distribution of the *nsubj* relation across the bins.

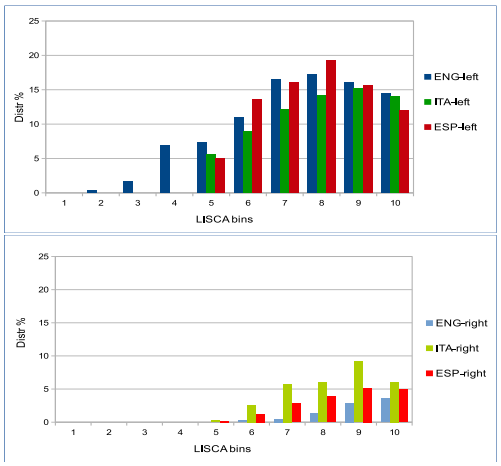

Figure 9: Direction of the *nsubj* relation across the bins.

by LISCA in the last bins. This in line with the Dependency Length Minimization principle (Temperly, 2007), i.e. the idea that languages tend to place closely related words close together since shortest dependency links reduce the human processing load and make the sentence comprehension process easier (Gibson, 1998). If we consider the average length of dependencies for each language, we note that they are quite similar, i.e. they are about three–token long on average. However, the LISCA ranking highlighted interesting differences between typologically different languages. The average length of Italian and Spanish dependencies share a more similar trend with respect to the English ones: the confidence interval of the difference between means is lower for Italian and Spanish (0.077 points) than both English/Italian (0.47 points), and English/Spanish (0.54 points). On the basis of this we can hypothesize an interesting interplay between dependency direction and dependency length: among the underlying properties of relations occurring in the last bins there are structural factors at work such as dependency length.

5.2. The case of *nsubj* and *amod* relations

Let us focus now on specific dependency relations, namely nominal subjects (*nsubj*) and adjectival modifiers (*amod*), and their distribution across the LISCA bins. These relations correspond to constructions widely investigated in the typological literature, and for this reason they represent two challenging testbeds for our methodology of analysis. Given the typology of features underlying LISCA, different factors contribute to the distribution of the same relation across the bins, concerning local features such as the linear ordering of words involved in the relation or their distance, to more global characteristics reflecting the position of the dependency arc within the overall dependency tree. In what follows, we try to reconstruct what are properties playing a role in determining the distribution of different word order patterns across the LISCA bins.

Nominal subjects (*nsubj*). Figure 8 reports the distribution of nominal subjects (both lexical and pronominal) across the LISCA bins. As it can be noted, the three languages are characterized by different distributions. First, whereas for English *nsubj* relations already appear in the top positions of the LISCA ranking, i.e. in the 2nd bin, the first occurrences of *nsubj* relations for Italian and Spanish are in the 4th and 5th bins respectively. Second, main differences can be observed at the level of frequency distributions. Both differences can be explained by considering that whereas Italian and Spanish are pro–drop languages English obligatorily requires an explicit subject. This is also reflected by the frequency distribution of pronominal subjects, which correspond to 59% of the total amount of nominal subjects in the English treebank, while they cover only about 3% of the cases in both Italian and Spanish treebanks.

Consider now the distribution of *nsubj* relations by dependency direction: in line with what reported in Section 2., we focus now on lexical (i.e. non-pronominal) subjects only. In Figure 1, it is shown that in all considered languages subjects are mostly right-headed, whereas significant differences are recorded for left–headed subjects whose percentage is much higher for Italian and Spanish (i.e. 30% and 18% respectively) than for English (9%). Figure 9 reports, for the three languages, the distribution of right–headed subjects (i.e. $d > h$) vs left–headed ones (i.e. $h < d$) across the LISCA bins. Left–headed subjects turned out to be mainly concentrated in the second half of the LISCA bins, starting from the 5th one. Consider as an example the following *nsubj* relation ranked in the 10th bin for Italian: 'Dalla rabbia dei valonesi non si salva *niente e nessuno*' (lit. 'From the rage of Valaisers not be saved *nothing and nobody*', 'From the rage of Valaisers nothing and nobody can be saved'). The frequency of left–headed subjects for the English language is much lower than for Romance languages, being they mainly restricted to parenthetical clauses, such as for example *[...], said Bush*, and existential clauses.

We have seen that the LISCA bins progressively gather occurrences of rarer and less prototypical relation instances, such as left-headed subjects, whose occurrence mainly concentrates in the last four bins. Yet, as already pointed out in Section 2., the position with respect to the governing head may be influenced by different linguistic properties. Consider the following examples for the three languages:

- Italian: '*La proposta presentata dalla commissione conformente al suo mandato* costituisce un punto di partenza' (lit. '*The proposal presented by the commission in accordance with its mandate* represents a starting point'), where the subject *proposta* ('proposal') is modified by a participial phrase which creates a long-distance *nsubj* dependency;

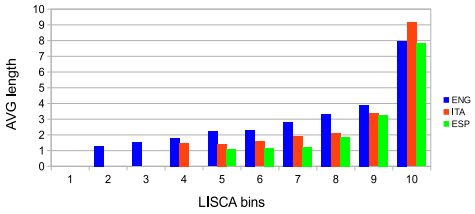

Figure 10: Average length of *nsubj* relations across the bins.

- Spanish: '*El descenso de la población indígena y la falta de mano de obra para los obrajes españoles* originó el comercio de pobladores secuestrados ...' (lit. '*The decline of the indigenous population and the lack of labor for the Spanish obrajes* led to the trade of kidnapped settlers ...'), where a coordinated subject determines a long-distance dependency link;

- English: 'Sergey Brin has actually a mathematical proof that the company's self–driven research *strategy, which gives employees one day a week to do research projects on their own, is a good, respectable* idea', where 22 tokens occur between the lexical subject *strategy* and its syntactic head.

All *nsubj* relations exemplified above have been ranked in the last (i.e. 10th) LISCA bin of each ordered list of dependencies: they all represent long distance dependencies involving a right-headed subject. These examples suggest an interesting and complex interplay between dependency length and the position of the subject with respect to the governing head in influencing word order patterns, which is worth being investigated.

In Section 5.1. we hypothesized a correlation between word order and dependency length. Let us now explore how the two interact for a specific dependency relation, *nsubj*. In Figure 10, it is reported that for all languages "heavier" (i.e. long distance) subjects (both right- and left-headed) are ranked in the last bins. For most part of the bins, Romance languages show stronger similarities with respect to English. Despite these differences, each language appears to follow the same trend, characterized by the fact that the ordered bins contain increasingly longer dependencies. Two questions arise at this juncture: i) whether dependency length is also influenced by the syntactic realization of the subject, i.e. lexical or pronominal; ii) for lexical subjects, whether and to what extent dependency length influences subject order patterns.

Concerning the first question, our intuition is that longer dependencies are typically represented by lexical subjects, i.e. by *nsubj* relations with a noun as dependent. This is confirmed by the fact that for all languages *nsubj* relations ranked in the first LISCA bins mostly have a pronoun as a dependent. On the other hand, relations ranked in the last bins are represented by long-distance dependencies with lexical subjects, as exemplified by the sample sentences reported above for the three languages.

Consider now the second open issue above, in particular the distribution of left– and right–headed subjects across the

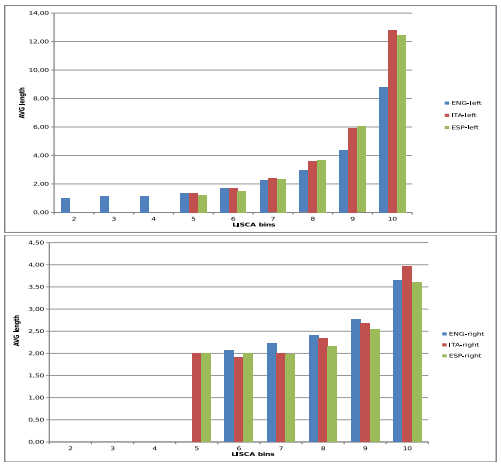

Figure 11: Average length of right– and left–headed lexical *nsubj* across the LISCA bins.

LISCA bins with respect to their length, to assess the correlation, if any, between dependency length and word order. As shown in Figure 11, for all languages the maximum average length of right–headed *nsubj* relations is higher (i.e. about 12 tokens for the Romance languages and about 8 for English) than the maximum average length of less frequent left–headed subjects (which is about 4 tokens for all languages). Besides reported differences among languages, a similar trend can be reported here: namely, all languages tend to minimize the dependency length when an alternative (with respect to the dominant, i.e. most frequent one) word order is used. Dependency length minimization varies across languages: whereas in English the average dependency length of right–headed subjects is only slighly lower with respect to left–headed ones (i.e. 2.62 vs 2.76), for Italian and Spanish length minimization is more evident (i.e. 2.48 vs 4.64 and 2.38 vs 4.54 respectively). From what seen so far, we can put forward the hypothesis that length minimization plays a stronger role with less frequent, and therefore more marked, word orders: in other words, if on the one hand marked order is associated with minimized dependency length, on the other hand the dominant unmarked order allows significantly longer dependencies.

Adjectival modifiers *amod*. Figure 12 reports the distribution of adjectival modifiers across the bins: it can be noticed that while for English they already appear in the 1st bin, for both Romance languages the first occurrences of *amod* relations start in the 2nd bin. This difference can be explained by considering that even in this area English is characterized by a generally fixed order or at most slightly variable structures, which are more easily predictable. Their distribution across the bins also varies significantly, with Italian and Spanish sharing a similar trend as opposed to English.

Consider now the relative order of Adjective and Noun: in Figure 2 it was shown that the pre–nominal adjectives are more common in the English UD treebank than in the treebanks of the Romance languages. The distribution of the *amod* relation across the LISCA bins shows that right–headed adjectives are more prototypical in English than in Italian and Spanish (see Figure 2). This is due to the fact

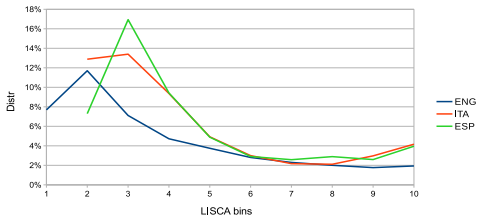

Figure 12: Distribution of the *amod* relation across the bins.

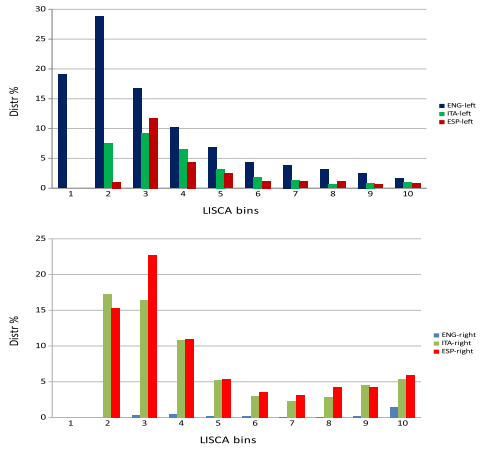

Figure 13: Direction of the *amod* relation.

that, differently from English, a marked position of adjectives is allowed in Italian and Spanish. The distribution of relations across the LISCA bins allows detecting adjectival modifiers occurring in prototypical constructions.

Differently from what observed for subjects, the average distance between d and h in *amod* relations remains tendentially constant through the bins (see Figure 14), ranging between 1 and about 3. Some differences can be observed if we compare English and the two Romance languages: the latter tend to be characterized by shorter relations. This may be explained in terms of adjacency to the nominal head: i.e. in the left–headed position adjectival modifiers are typically adjacent to the head, which is not necessarily the case in the case of prenominal position where the adjectival modifier can be separated from the head by elements that belong to the same subtree.

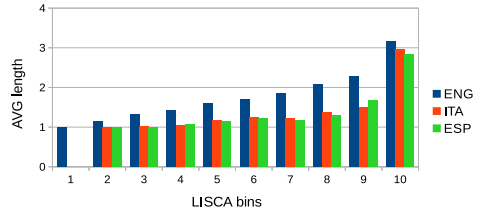

Figure 14: Average length of *amod* relations.

6. Conclusion

In this paper we presented a new methodology aimed at acquiring typological evidence from "gold" treebanks for different languages. In particular, we investigated whether algorithms for measuring the plausibility of dependency relations within the output of dependency parsers could be exploited to acquire quantitative evidence from gold treebanks to shed new light in linguistic typological studies. The methodology was tested in a case study carried out on UD treebanks for different languages: English, Italian and Spanish. By relying on a wide range of linguistic properties aimed at weighting the plausibility of a given dependency arc, it has been possible to reconstruct an articulated profile of word order patterns attested in the languages considered, in line with the literature and which has been enriched with new interesting insights. Starting from the study of general word order trends and their relationship with dependency length, we focused on two dependency relations widely investigated in the typological literature, nominal subjects (*nsubj*) and adjectival modifiers (*amod*). For both of them, word order similarities and differences were reported for all languages, significantly enriching the picture emerging from typological databases and showing the added value of the LISCA-based methodology with respect to simple frequency distributions. We also investigated the underlying properties influencing the preference towards one word order or the other. Among them, dependency length turned out to play a significant role: its impact, however, appears to vary according to whether a dominant or marked order is used. Dependency length minimization seems to be at work with less frequent order patterns, thus suggesting an interesting interaction between word order and dependency length.

However, the potentialities of the method are not restricted to the area of typological studies. Nowadays, linguistic typology is starting to play a role in multilingual Natural Language Processing (O'Horan et al., 2016). While the growing importance of typological information in supporting multilingual tasks has been recognized, existing typological databases such as WALS have still a partial coverage, and most importantly here, do not always reflect real language use. Methods for automatic induction of typological information are still at the beginning: this paper represents a promising attempt in this area. It is a widely acknowledged fact that word-order affects the automatic analysis of sentences: free–order languages are harder to parse (Gulordava and Merlo, 2015c). Acquired information from real language usage can be used among the "selective sharing parameters" in a cross–lingual transfer parsing scenario (Naseem et al., 2012).

Further directions of research include: the application of the methodology to other languages, including typologically distant ones, to reconstruct shades of typological proximity starting from real language usage; the analysis of other dependency relations as well as of more complex structures such as dependency subtrees; the extension of the range of properties expected to influence the preference towards a given word order pattern. Last but not least, we are planning to test the effectiveness of acquired typological information in a cross–lingual parsing scenario.

Acknowledgements

The work reported in the paper was partially supported by the 2–year project (2017-2019) *UBIMOL, UBIquitous Massive Open Learning*, funded by Regione Toscana (BANDO POR FESR 2014-2020).

7. Bibliographical References

Alzetta, C., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2018). Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 201–210, Prague, Czech Republic, January.

Baldwin, T. and Kordoni, V., (2011). *Interaction between Linguistics and Computational Linguistics*. CSLI Publications.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August.

Che, W., Guo, J., and Liu, T. (2014). Reliable dependency arc recognition. *Expert Systems with Applications*, 41(4):1716–1722.

Dell'Orletta, F., Venturi, G., and Montemagni, S. (2011). ULISSE: an unsupervised algorithm for detecting reliable dependency parses. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, pages 115–124, Portland, Oregon, USA, June. Association for Computational Linguistics Stroudsburg, PA, USA.

Dell'Orletta, F., Venturi, G., and Montemagni, S. (2013). Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas*, 2:125–136.

Dell'Orletta, F. (2009). Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics, DepLing 2015, August 24-26 2015, Uppsala University, Uppsala, Sweden*, pages 91–100.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113.

Gulordava, K. and Merlo, P. (2015a). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of latin and ancient greek. In *Proceedings of the Third International Conference on Dependency Linguistics, DepLing 2015, August 24-26 2015, Uppsala University, Uppsala, Sweden*, pages 121–130.

Gulordava, K. and Merlo, P. (2015b). Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 247–257.

Gulordava, K. and Merlo, P. (2015c). Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of Conference on Computational Natural Language Learning (CoNLL)*, pages 247–257, Beijing, China, July.

Haspelmath, M. (2016). Against Markedness (and what to Replace it with). *Journal of Linguistics*, 42:25–70.

Kay, M. (2005). Acl lifetime achievement award: A life of language. *Computational Linguistics*, 31(4).

Kučera, H. (1982). Markedness and frequency: a computational analysis. In *Proceedings of COLING 82*.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Castelló, N. B., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Merlo, P. (2016). Quantitative computational syntax: some initial result. *Italian Journal of Computational Linguistics*, 2.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju, Republic of Korea, July.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing - Proceedings of the 16th International Conference, CICLing 2015, Part I*, pages 3–16, Cairo, Egypt, April.

O'Horan, H., Berzak, Y., Vulic, I., Reichart, R., and Korhonen, A. (2016). Survey on the use of typological information in natural language processing. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1297–1308.

Reichart, R. and Rappoport, A. (2009). Sample selection for statistical parsers: Cognitively driven algorithms and evaluation measures. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2009)*, pages 3–11. Association for Computational Linguistics Stroudsburg, PA, USA.

Silveira, N., Dozat, T., de Marneffe, M., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Straka, M., Hajic, J., and Strakova, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Temperly, D. (2007). Minimization of dependency length

in written english. *Cognition*, 105(2):300–333.

Tusa, E., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2016). Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessitá. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, pages 3–16, Napoli, Italy, December.

8. Language Resource References

J. Nivre and A. Željko and A.Lars and et alii. (2017). *Universal Dependencies 2.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, 2.0, ISLRN http://hdl.handle.net/11234/1-1983.

M. S. Dryer and M. Haspelmath. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, ISLRN http://wals.info/.

M. S. Dryer. (2013a). *Order of Adjective and Noun*. Max Planck Institute for Evolutionary Anthropology, ISLRN http://wals.info/chapter/87.

M. S. Dryer. (2013b). *Order of Subject and Verb*. Max Planck Institute for Evolutionary Anthropology, ISLRN http://wals.info/chapter/82.

SSWL. (2009). *Syntactic Structures of the World's Languages*. ISLRN http://sswl.railsplayground.net/.