Linguistically-driven Selection of Difficult-to-Parse Dependency Structures

Chiara Alzetta*
DIBRIS, Università degli Studi di Genova;
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab

Felice Dell'Orletta**
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab

Simonetta Montemagni†
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab

Giulia Venturi‡
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab

The paper illustrates a novel methodology meeting a twofold goal, namely quantifying the reliability of automatically generated dependency relations without using gold data on the one hand, and identifying which are the linguistic constructions negatively affecting the parser performance on the other hand. These represent objectives typically investigated in different lines of research, with different methods and techniques. Our methodology, at the crossroads of these perspectives, allows not only to quantify the parsing reliability of individual dependency types, but also to identify and weight the contextual properties making relation instances more or less difficult to parse. The proposed methodology was tested in two different and complementary experiments, aimed at assessing the degree of parsing difficulty across (a) different dependency relation types, and (b) different instances of the same relation. The results show that the proposed methodology is able to identify difficult-to-parse dependency relations without relying on gold data and by taking into account a variety of intertwined linguistic factors. These findings pave the way to novel applications of the methodology, both in the direction of defining new evaluation metrics based purely on automatically parsed data and towards the automatic creation of challenge sets.

1. Introduction and Motivation

The analysis of dependency parsing performance has long attracted the attention of many studies mostly devoted to assess approaches to quantitatively measure parsing

* DIBRIS, Università degli Studi di Genova; Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab E-mail: chiara.alzetta@edu.unige.it
** Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab E-mail: felice.dellorletta@ilc.cnr.it
† Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab E-mail: simonetta.montemagni@ilc.cnr.it
‡ Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab E-mail: giulia.venturi@ilc.cnr.it

© 2020 Associazione Italiana di Linguistica Computazionale

accuracy and to identify the variety of factors making specific constructions particularly difficult-to-parse. To this aim, there are several efforts addressed towards the definition of evaluation metrics able to assess errors of analysis adequately, namely without biases originating in language-specific structural characteristics.

In this respect, the Universal Dependencies (UD) initiative (Nivre 2015), aimed at developing cross-linguistically consistent treebank annotation for many languages, represents a milestone: it has created the premises for the introduction of evaluation metrics that guarantee the comparability of the quality of automatically produced parses across typologically different languages. This is the case of the metrics introduced for the 2017 and 2018 CoNLL shared tasks on Universal Dependency parsing (Zeman et al. 2017, 2018) which have been specifically and explicitly devised to score parsing systems avoiding biases originating in analytic vs. synthetic languages (Nivre and Fang 2017). To maximise comparability across languages, these new metrics are mainly focused on dependencies holding between content words and, similarly to those devised previously, they are aimed at quantitatively assess the accuracy of a parser in terms of overall correctness, which is computed against gold labelled data sets.

However, the broad perspective sketched above does not account for the linguistic constructions that, being difficult-to-parse, may negatively affect the parser performance. This is rather the focus of a second and complementary line of research, which is aimed at devising methods and techniques to identify and weight the factors negatively affecting the performance of a parser. We briefly mention here the main approaches proposed to investigate the linguistic sources of parsing complexity. Far from being an state-of-the-art survey of the literature on the topic, this short overview is meant to let the reader appreciate the main contribution of our paper in this area of research. A quite detailed comparative analysis of errors made by different types of dependency parsers with respect to more general and structural properties of the input is carried out, for example, by (McDonald and Nivre 2007). They reviewed a set of length and graph factors that resulted to negatively affect parsing accuracy, with the former factors being concerned with sentence and dependency length, and the latter with characteristics such as the distance to the root node, or the number of children or siblings or non-projectivity. Similar investigations are performed by (Rimell, Clark, and Steedman 2009), who evaluated the performance of different parsers against a test suite of unbounded dependency constructions, or by (Droganova et al. 2018), whose study is aimed at isolating parsing errors in the analysis of elliptical constructions. These and similar studies share a common methodology: they start from a set of a-priori known constructions that result challenging for many existing state-of-the art parsers and whose parsing complexity is evaluated on the basis of available manually revised data. Among these constructions, dependency distance is, for example, one of the most explored factors (McDonald and Nivre 2011; Gulordava and Merlo 2015; Merlo 2015), which makes sentences particularly hard to parse in free-word order languages (Gulordava and Merlo 2016). In this situation, it seems that there are research questions which remain unanswered at the moment, at least to our knowledge: for instance, whether linguistic properties such as dependency length affect to the same extent the analysis of all dependency types; or, if this is not the case, what are the structural properties determining the parsing difficulty of different instances of the same dependency type.

Starting from these premises, in this paper, we introduce a methodology which combines the two lines of research sketched above. As discussed in the following paragraphs, the proposed approach operates at two different levels, originating in distinct but related goals. On the one hand, a macro-level analysis of the parsing output is carried out with the aim of quantifying the reliability of the different types of automati-

cally generated dependency relations without using gold data, as opposed to standard evaluation metrics. On the other hand, it operates at the micro-level of relation instances to identify, for specific relation types, which are the linguistic constructions negatively affecting the parser performance. By combining the two perspectives, the methodology allows not only to quantify the parsing reliability of individual dependency types (or subsets of them), but also to rank relation instances by parsing difficulty. In this way, it becomes possible to simultaneously single out *i)* easy-to-parse dependencies that become difficult, possibly due to the higher complexity of their context of occurrence, or, the other way round, *ii)* difficult-to-parse ones that become easy-to-parse when occurring in prototypical linguistic constructions.

This combined approach has a twofold goal, a short- and a long-term one. The short-term goal consists in the **identification and weighting of sources of parsing complexity** which are detected without relying on gold data (not always available for testing parsing accuracy). The long-term goal concerns the **construction of challenge sets containing instances of difficult-to-parse structures** extracted from corpora that can thus be used to test or even re-train parsers with a view to improving their performance. In this paper, we focus on the first goal. In order to meet this goal, we rely on LISCA (*LInguiStically–driven Selection of Correct Arcs*) (Dell'Orletta, Venturi, and Montemagni 2013), an algorithm developed to measure the reliability of automatically generated dependency relations simultaneously taking into account a wide range of factors also including the linguistic context in which they occur. The score returned by LISCA quantifies the reliability of individual dependency relations and can be used to rank dependencies accordingly.

In particular, in this paper we try to answer two complementary and intertwined research questions:

- **RQ1.** is it possible to identify and weight the parsing reliability of automatically generated dependency relations without resorting to gold data?
- **RQ2.** what are the underlying properties making specific dependency relation instances easy- or difficult-to-parse?

The rest of the paper is organised as follows: in Section 2, we introduce the overall method together with the LISCA algorithm. In Section 3 we introduce our experiments and the considered corpora, representative of three Indo-European languages belonging to two different typological genera all belonging to the Indoeuropean language family, i.e. a Germanic language (English, ENG) and two Romance languages (Italian, ITA, and Spanish, SPA). Sections 4 and 5 present the investigations aimed at exploring, respectively, parsing difficulty over dependency relations types and individual instances. Section 6 highlights the contribution of the paper from both methodological and linguistic perspectives, with a specific view to its potentialities in different scenarios.

2. Method

To pursue the goal of identifying difficult-to-parse dependency relations in annotated corpora, we defined the methodology sketched in Figure 1, which is organised as follows:

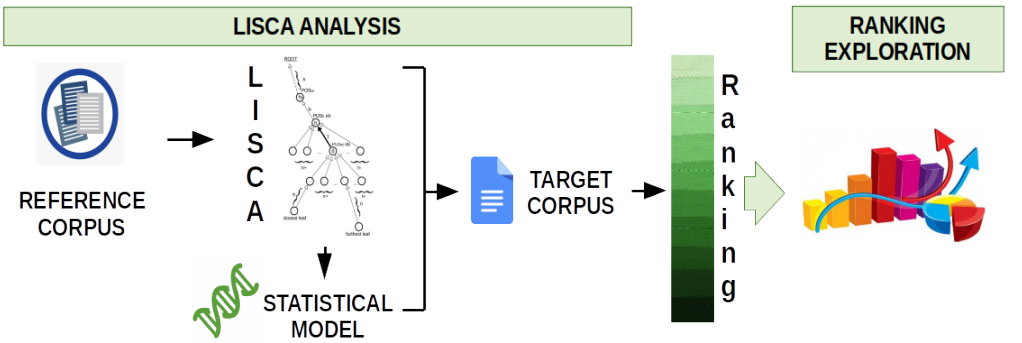


Figure 1
Method work–flow.

- **LISCA analysis phase**, which consists in the assignment of a *plausibility score*, i.e. a score expressing the reliability of the analysis, to each dependency relation occurring in an annotated corpus, on the basis of both the relation type and its linguistic context;
- **Ranking exploration phase**, which is devoted to the analysis of the list of dependency relations in the annotated corpus ranked by plausibility, with a specific view to the distribution of relation types and the linguistic context where specific relation instances occur.

The plausibility score assigned to each dependency relation in the corpus under analysis has been computed by relying on LISCA. As detailed in Section 2.1, this algorithm is used to obtain a Statistical Model containing statistics that concern the distribution of a wide range of linguistic properties of the dependency relations, which we will refer to as 'features' in the remainder of the paper. Such features are automatically extracted from a large corpus of automatically parsed sentences (henceforth, the *Reference Corpus*). The resulting Statistical Model is then exploited to assign a score to a target dependency relation, defined as a triple d(ependent), h(ead), and t(ype) of dependency linking d to h (hereafter referred to as *DR*). Hence,

$$DR = (d, h, t)$$

The LISCA–based ranking of DRs in the corpus being analysed (henceforth, *Target Corpus*) is then explored to identify and discern easy-to-parse vs difficult-to-parse DR types and constructions, as described in Section 2.2.

2.1 The LISCA Algorithm

LISCA is an unsupervised algorithm aimed at assigning a *plausibility* score to each DR in a Target Corpus based on the statistics acquired from a Reference Corpus. The algorithm operates in two steps:

1. collection of statistics about a set of linguistically motivated features extracted from the Reference Corpus (RC) to build a Statistical Model of the language;

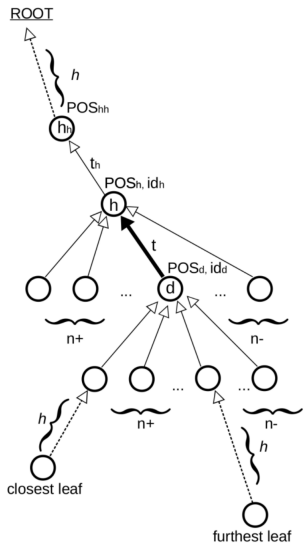


Figure 2
Features computed by LISCA for each dependency relation in the corpus.

2. assignment of a plausibility score to each DR in the Target Corpus (TC) on the basis of the Statistical Model built at the previous step.

Extending a metaphor borrowed from the linguistic literature originally introduced by (Jakobson 1973), we look at the Statistical Model obtained with LISCA as encoding the DNA of the language being analysed. The features (detailed in Section 2.1.1) considered by the algorithm to build the statistical model cover, for each DR instance, both global and local properties of the parsed sentences where DRs occur. In fact, by using a large set of monolingual examples as Reference Corpus, we can look at the obtained statistical model as a good approximation of the distribution of linguistic phenomena in a given language. The LISCA score can thus be seen as reflecting the degree of similarity of the linguistic contexts where DRs occur in the Target Corpus with respect to the statistics acquired from the Reference Corpus.

For the specific concerns of this study, we used LISCA in its delexicalised version: this allows to abstract away from variations resulting from lexical effects that may affect the cross-lingual comparability of the obtained results.

2.1.1 Features

The features underlying LISCA are aimed at characterising each dependency relation appearing in RC with respect to local and global properties of the dependency tree, referring respectively to the linear ordering and the hierarchical structure of the sentence (see Fig. 2). They include:

- the linear distance of d (i.e. length of the shortest path) from *i)* the root of the dependency tree, *ii)* the closest leaf node, and *iii)* the most distant leaf node;
- the number of "siblings" of d, distinguishing between those located to its right and/or left in the linear sequence of words of the sentence;

- the number of "children" nodes of d, again partitioned into two classes following their liner ordering with respect to d.
- dependency length, i.e. the distance in terms of intervening tokens between d and h;
- dependency direction, used to distinguish between head-initial and head-final dependency arcs;
- *ArcPOSFeat*, a complex feature computing the associative strength between d and h and d and h_h (the head of the head governor), for which we provide more details below.

These features are said to be "linguistically-motivated" since they are based on the dependency tree structure, and in particular they are focused on structures widely agreed in the literature to reflect the syntactic and parsing complexity of sentences. Such set of features represents the dependency-based counterpart of the features underlying syntactic complexity measures, such as node-counting algorithms that count the number of nodes in the phrase markers of syntactic constructions: this is the case of, e.g., local nonterminal count (Frazier 1985) or the depth algorithm (Hawkins 1994), as well as of word-counting algorithms based on ratios involving the length of constituents in terms of words (Yngve 1960).

2.1.2 LISCA Score

The Statistical Model built on the basis of the features listed in Section 2.1.1 is used to compute the LISCA score which is associated with each dependency relation DR appearing in the Target Corpus TC. To this aim, we refer to $LISCA(TC_i)$ as the LISCA score computed for the ith dependency relation in TC. Specifically, as described in (Dell'Orletta, Venturi, and Montemagni 2013), the LISCA score is computed as a simple product of the weights associated to the features listed above. More in detail, $LISCA(TC_i)$ is computed as follows:

$$LISCA(TC_i) = \prod_{y=1}^{n} Weight(TC_i, f_y, L(S), r, C)$$

In the above, f_y is the yth feature of TC_i; $L(S)$ is the length (in terms of tokens) of the sentence including TC_i; the parameter r refers to a numerical value[1]; C refers to different configurations where the value of f_y is computed, as detailed below.

The function $Weight(TC_i, f_y, L(S), r, C)$ computes the weight of each feature f_y, i.e. the probability that f_y assumes a given value when the yth feature occurs in a specific context of TC under specific conditions C. Specifically, the function is defined as:

$$Weight(TC_i, f_y, L(S), r, C) = \prod_{\forall c \in C_{f_y}} \left(\frac{F(V(f_y), Range(L(S), r), c)}{|Range(L(S), r), c|}\right).$$

The function $F(x, y, z)$ computes the frequency of x in TC as conditioned by y and z. Among the parameters of function F, $V(f_y)$ refers to the value of f_y; $Range(L(S), r)$

1 In all the experiments reported in this work, we set r=2.

defines the sentence length range covering values from $L(S) - r$ to $L(S) + r$: in other words, it restricts the considered subsets of sentences to those having a sentence length falling into the defined range; c defines multiple conditions for computing $V(f_y)$. Specifically $C = \langle N, P, D \rangle$, thus c may vary with respect to 3 different configurations aimed at defining different subsets of relations. When $c = N$, we compute the probability of observing a certain value of f_y within all the relations of TC appearing in sentences whose length falls into the defined range. When $c = P$, we compute how many times f_y assumes a given value among relations sharing the same POS (POS_d); while when $c = D$ we consider only links sharing the same DR type t. Accounting for such configurations allows to compute the probability of a certain feature value both as a relative frequency over the entire corpus and as conditioned by morpho-syntactic and syntactic characteristics of the dependency tree. Finally, $|Range(L(S), r), c|$ is the number of relations in TC appearing in sentences whose length is in the range $Range(L(S_i), r)$ and meeting the condition c.

All features weights are computed as described above, apart from the complex feature *ArcPOSFeat*, whose weight is defined as follows:

$$Weight(TC_i, ArcPOSFeat, L(S), r, C) =$$
$$\frac{F((POS_d, POS_h, t))}{\sum_X F((POS_d, X, t))} \cdot \frac{F((POS_d, POS_h, t))}{\sum_X F((X, POS_h, t))} \cdot \frac{F(((POS_d, POS_h, t)(POS_h, POS_{hh}, t_h)))}{F((POS_d, POS_h, t))} \cdot$$
$$\cdot \frac{F(((POS_d, POS_h, t)(POS_h, POS_{hh}, t_h)))}{F((POS_h, POS_{hh}, t_h))} \cdot \frac{F(((POS_d, POS_h, t)(POS_h, POS_{hh}, t_h)))}{\sum_X F((((POS_d, X, t))(X, POS_{hh}, t_h)))}$$

The triple (POS_d, POS_h, t) is the relation TC_i and $F((POS_d, POS_h, t))$ is its frequency in TC; X is a variable, thus $\sum_X F((POS_d, X, t))$ indicates that we take into account the sum of frequencies over any h of d linked by t occurring in TC; $((POS_d, POS_h, t)(POS_h, POS_{hh}, t_h))$ represents the sequence of two consecutive arcs going from d to the father of h (i.e., h_h) in TC_i within the dependency tree describing S.

Given the typology of properties considered when computing the LISCA score, it can be affirmed that it is both context-sensitive and frequency-based, i.e. it reflects the frequency of occurrence of DR types in actual language use as well as their occurrence in specific contexts.

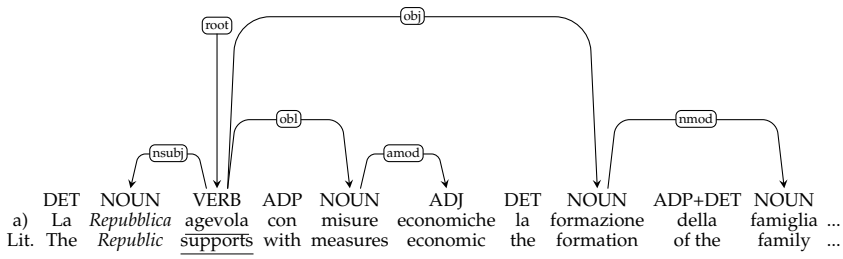
2.2 Exploration of the LISCA-based DR Ranking

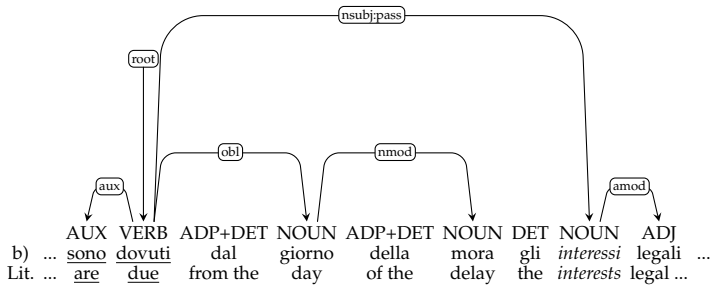
The LISCA score was used to rank the relations in the Target Corpus by decreasing score. The resulting ranking represents the starting point of the analyses carried out in this study: it has been explored with the final goal of investigating how DR types and the linguistic structures in which they occur are distributed in the ranked DR list.

In Section 2.1.2, we showed that the LISCA score is a context-sensitive and frequency-based measure. On the one hand, the measure is sensitive to changes in the context of occurrence of each specific DR since it reflects the degree of similarity of the linguistic context in which a given dependency relation occurs in the Reference and Target corpora. On the other hand, the score is based on the assumption that more frequently occurring syntactic structures are more likely to be correct than less frequent ones. From this, it follows that *higher LISCA scores* are assigned to DRs associated with linguistic contexts more frequently occurring in the Reference corpus; on the contrary,

lower scores identify DR instances occurring in less typical contexts, less common in the Reference corpus.

To better appreciate how the linguistic context where a DR instance occurs impacts on the LISCA score, consider the following examples showing different instances of Italian and English `nsubj` and Spanish `amod` relations, appearing in the corpora used for the present study[2].

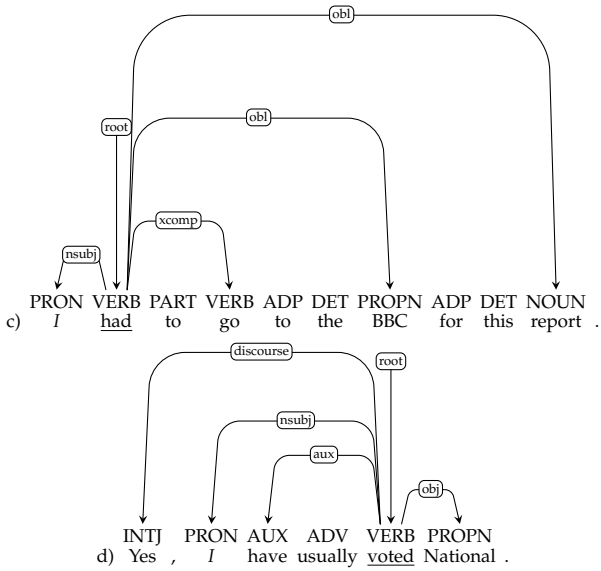




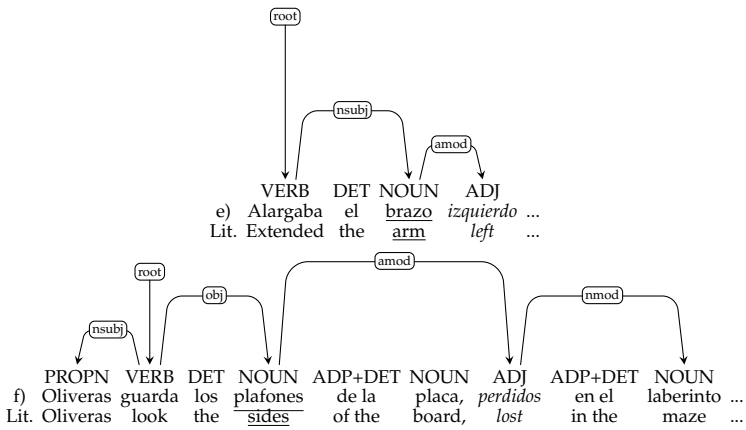
The subjects displayed in a), 'La *Repubblica* agevola' (lit. 'The *Republic* supports'), and b), 'Sono dovuti [...] gli *interessi*' (lit. 'Are due [...] the *interests*'), represent two different examples of Italian subjects. In a), the subject immediately precedes the verb, which represents the typical Italian subject; on the other hand in b) we observe a less common example, with a long–distance passive nominal subject occurring in post-verbal position. The LISCA ranking reflects the frequency of the Italian subject constructions: a) that contains a short distance pre–verbal subject is more highly ranked than b) that, on the contrary, contains a less typical post–verbal and long-distant subject.

Consider now the two sentences in c) and d), exemplifying different cases of English pronominal subjects. Again, the plausibility score assigned to them by LISCA reflects the typicality and the frequency of the two constructions in the English language: c) occurs higher in the ranking with respect to d). In fact, while c) shows a SUBJ-VERB ordering involving a pronoun and a verbal root appearing in a short sentence, d) represents a less typical situation, with an interjection involved in a `discourse` relation. Additionally, by considering the length of the relation, we notice that the presence of elements intervening between the pronoun and its verbal head results in a longer subject relation in d) with respect to c).

2 In the examples, the dependent is italicised and the head underlined.



Moving to cases involving adjectival modifiers, example e), 'el brazo *izquierdo*' (lit. 'The arm *left*'), represents the typical post–nominal construction for Spanish, where the adjective immediately follows the modified noun. The example in f), 'los plafones [...] *perdidos*' (lit. 'the sides [...] *lost*'), is assigned a lower score by LISCA: despite the shared post–nominal position, we are in front of a longer `amod` relation.



The three pairs of examples reported above show that the ordering of DRs on the basis of the LISCA score reflects the degree of typicality of the linguistic contexts where they occur.

LISCA has been successfully applied in different scenarios, against both the output of dependency parsers and gold treebanks. As pointed out above, the score returned by LISCA was originally meant to identify unreliable automatically produced dependency relations (Dell'Orletta, Venturi, and Montemagni 2013). However, it has also been successfully used against "gold" dependency treebanks, in order to detect shades of syntactic markedness of syntactic constructions from a monolingual perspective (Tusa et al. 2016), or to acquire quantitative typological evidence from a multilingual perspective (Alzetta et al. 2018, 2019, 2020). Last but not least, it was also exploited to

identify anomalous annotations (going from annotation inconsistencies to real errors) from a monolingual perspective in "gold" treebanks (Alzetta et al. 2017).

Past uses of LISCA motivate and provide the foundations of the work presented here. For the specific purpose of this study, we are combining the two application perspectives. As illustrated in Section 3, the LISCA model has been built considering the same Reference Corpus, here considered as encoding the *DNA* of the languages taken into account, but has been applied against two versions of the same Target Corpus, i.e. the automatically parsed and manually revised ("gold") versions. The choice follows from the two-fold goal of both discerning easy-to-parse from difficult-to-parse dependency relations and of identifying associated linguistic constructions.

3. Experiments and Data

To address the research questions introduced in Section 1, two sets of experiments have been carried out, aimed at assessing the degree of parsing difficulty across *(a)* different DR *types*, i.e. the subsets of DR instances sharing the same syntactic label, and *(b)* different *instances* of the same DR type. In broad terms, the two types of analyses proposed in what follows respectively operate at the macro-level of DR-centred dependency evaluation metrics (a), and at the micro-level of specific linguistic constructions negatively affecting the parser performance with respect to specific DR instances (b).

In particular, Section 4 is devoted to compare and measure the correlation between the LISCA-based ranking of DR types and the rankings resulting from other measures (e.g. parsing evaluation measures). The analysis illustrated in Section 5 goes in more detail and investigates which are the global and local dependency tree properties that make specific instances of a DR type more or less difficult-to-parse. While the first experiment is carried out for all languages and all DR types, the second one is focused on one single DR type, i.e. nominal subject, which is analysed for Italian and English, selected for their higher typological distance among the languages taken into account.

Note that, from now on, we will use the following notation: `DR` identifies a specific type of dependency relation, extensionally corresponding to the set of relations sharing the same syntactic label, e.g. all instances of nominal subjects will be referred to as `NSUBJ`, objects as `OBJ`, determiners as `DET` etc.; a `DR instance`, corresponding to a specific dependency link in the annotated text, is referred to as e.g. `nsubj` for a specific instance of the `NSUBJ` type or `det` for `DET`.

3.1 Corpora and Languages

For each language taken into account (i.e., English, Italian and Spanish), two linguistically annotated corpora have been used: namely, a Reference Corpus for building the LISCA statistical model and a Target Corpus. The reference corpora used to collect the statistics to build the models represent a portion of the English, Italian and Spanish Wikipedia, for a total of around 40 million tokens for each language: this constitutes a set of examples large enough to reflect the actual distribution of phenomena in each language. As target corpora, we took the test sets of the *Universal Dependencies* (UD) treebank (Nivre et al. 2020) of each language in order to guarantee multilingual comparability thanks to the shared inventory of relations and annotation guidelines. Specifically, the following UD treebank test sets (released in July 2018, version 2.2) were used:

- English Web Treebank test set (25,095 tokens and 2,077 sentences) (Silveira et al. 2014);
- Italian Stanford Dependency Treebank test set (9,680 tokens and 482 sentences) (Bosco, Montemagni, and Simi 2013);
- Spanish Ancora UD treebank test set (52,617 tokens and 1,721 sentences) (Alonso and Zeman 2016).

Reference and target corpora were morpho-syntactically annotated and dependency parsed by the UDPipe pipeline (Straka, Hajič, and Straková 2016) trained on the Universal Dependency treebanks, version 2.2 (Nivre et al. 2020)[3].

4. Dependency-type Analysis

This first experiment is aimed at investigating whether and to which extent the DR ranking based on the LISCA score is aligned with the rankings based on established dependency-based parsing evaluation metrics as well as on characteristics widely acknowledged to affect parsing complexity such as dependency length (McDonald and Nivre 2007).

4.1 Experimental Setting

To address the question of whether the LISCA score can be interpreted as a metric reflecting the degree of parsing complexity associated with individual dependencies, we ranked the DR types in the Target Corpus of each language on the basis of different metrics, listed below:

- **LISCA Score**: the LISCA score assigned to each DR type, computed as the arithmetic mean of all the scores assigned to each DR instance sharing the same label;
- **F-score**: the labelled F-score computed for each DR type by taking into account both head and dependency label assignments, obtained using the evaluation script released for the 2018 CoNLL shared task[4] which was modified to return dependency-specific scores;
- **Link Length**: for each DR type, the average dependency length, computed as an arithmetic mean of the number of tokens occurring between *d* and *h*, of all DR instances sharing the same label.

For each language, the LISCA scores were computed twice, namely against (i) the gold test set of the UD treebank, and (ii) the same test set automatically parsed with UDPipe (pre–trained using the treebank version 2.2): they will be henceforth referred to as *LISCA Score Gold* and *LISCA Score Parsed* respectively. This choice was aimed at assessing whether and to what extent the *LISCA score* applied to automatically generated DRs could be seen as a reliable dependency–based measure of parsing difficulty. Note

3 Note that at the writing time the subsequent versions of the UD treebanks used in these experiments didn't involve major revisions with respect to version 2.2 but only refinements and error fixing.
4 `http://universaldependencies.org/conll18/conll18\_ud\_eval.py`

that the *Link Length* measure was computed only with respect to the gold version of the Target Corpus.

In order to guarantee cross-language comparability of the results, this experiment has been carried out on the subset of 24 DR types shared by the test sets of the three languages taken into account. For each metric, a ranking of DR types was created. In the case of the *F-score*, DR types were ranked by decreasing scores, with DRs that are more likely to be parsed accurately in the top positions and, conversely, DRs more difficult to be parsed at the bottom. Similarly, the *LISCA scores*, applied to both parsed and gold dependency annotations, are ordered from higher to lower, with more or less reliable dependency relations in higher / lower positions respectively. Finally, the *Link Length* ordering contains DRs characterised by shorter links on the top part of the ranking, and longer links at the bottom.

In what follows, we report, for each DR type, the position in the ranking based on different orderings, rather than the values associated for each metric: this makes it easier to compare the rankings resulting from the different metrics within the same language and across different languages. The four rankings are expected to reflect parsing difficulty from different perspectives. By comparing the rank position of the same DR type as well as the relative ordering of DRs across different metrics for the same language and across languages, we will investigate to which extent the different measures correlate. In particular, we are interested in *i)* assessing whether the LISCA scores can be seen as reliable proxies of parsing difficulty of individual relation types, and *ii)* detecting similarities and differences between the rankings obtained with the *LISCA Score Parsed* vs *LISCA Score Gold*.

4.2 Results

The tables in Figure 3 show, for each language, the 24 shared DR types, ranked on the basis of the different metrics considered in this experiment and their correlations.

Each ranked list of DRs is graphically represented as a gradient of different shades of green, with lighter colours representing higher positions and darker shades as we move down in the ranking. For each language, the list of DRs (*DEPREL* column in Figure 3) is ranked by decreasing F-scores: since this measure is widely used as reference parsing evaluation metric, we adopt the F-score ranking as a benchmark for identifying the DR types that generally obtain more (or less) accurate parsing results. As an example, consider `DET` and `ADVCL` DR types, respectively corresponding to determiners and adverbial clauses. Determiners appear in the first or second position in the F-score ranking for all the three languages, meaning that the F-score obtained by this relation is generally high, i.e. most `DET` instances are correctly parsed. On the other hand, we find the `ADVCL` relation in the bottom part of the ranking (i.e. in 22nd, 23rd and 20th positions for Italian, Spanish and English respectively), indicating that the parser has more difficulty in providing the correct analysis for this DR type for all considered languages.

At a first glance, Figure 3 offers interesting insights about difficult-to-parse DRs. By comparing the rankings resulting from all metrics, different results immediately stand out. First, the proposed visualisation, in fact, shows quite clearly that the rankings based on parsed and gold LISCA scores are almost identical. Such an impressionistic observation is confirmed by the Spearman's correlation values between rankings reported for each language in the correlation matrices in Figure 3: *LISCA Score Parsed* and *LISCA Score Gold* rankings show an extremely high correlation, namely 0.99 for all languages. Second, the LISCA rankings and the DR list based on the F-score also show

ITALIAN

DEPRELS	*F-score Rank*	*LISCA Parsed*	*LISCA Gold*	*Link Length*
case	1	4	4	7
det	2	2	2	3
aux	3	1	1	6
cc	4	7	7	16
amod	5	8	8	4
flat	6	9	9	2
mark	7	3	3	14
fixed	8	5	5	1
nsubj	9	14	14	18
advmod	10	10	11	11
cop	11	6	6	8
nummod	12	11	10	13
obj	13	16	17	10
nmod	14	15	15	15
punct	15	13	12	23
obl	16	17	16	20
iobj	17	12	13	5
acl	18	20	20	17
conj	19	18	19	22
xcomp	20	19	18	12
ccomp	21	21	21	21
advcl	22	22	22	24
appos	23	23	23	19
csubj	24	24	24	9

	F-score Rank	*Link Length*	*LISCA Gold*	*LISCA Parsed*
F-score Rank	1			
Link Length	0,49	1		
LISCA Gold	0,87	0,61	1	
LISCA Parsed	0,88	0,62	0,99	1

SPANISH

DEPRELS	*F-score Rank*	*LISCA Parsed*	*LISCA Gold*	*Link Length*
det	1	1	1	1
case	2	2	2	6
aux	3	5	5	3
flat	4	10	10	7
amod	5	7	7	5
cc	6	3	3	11
mark	7	4	4	15
cop	8	6	6	8
nummod	9	9	9	4
nsubj	10	14	14	17
obj	11	15	15	10
punct	12	11	11	24
advmod	13	13	13	9
fixed	14	8	8	2
nmod	15	16	16	13
xcomp	16	22	21	14
ccomp	17	23	23	23
obl	18	18	19	18
acl	19	21	22	19
csubj	20	24	24	20
conj	21	19	18	21
appos	22	17	17	16
advcl	23	20	20	24
iobj	24	12	12	12

	F-score Rank	*Link Length*	*LISCA Gold*	*LISCA Parsed*
F-score Rank	1			
Link Length	0,58	1		
LISCA Gold	0,75	0,74	1	
LISCA Parsed	0,75	0,74	0,99	1

ENGLISH

DEPRELS	*F-score Rank*	*LISCA Parsed*	*LISCA Gold*	*Link Length*
aux	1	2	2	6
det	2	5	5	7
mark	3	1	1	13
case	4	3	4	9
nsubj	5	13	12	15
cc	6	4	3	14
obj	7	16	16	11
cop	8	7	7	10
amod	9	6	6	5
fixed	10	8	9	1
iobj	11	14	14	2
xcomp	12	20	20	12
advmod	13	10	10	8
nmod	14	15	15	16
punct	15	12	13	24
obl	16	17	18	19
ccomp	17	24	23	20
nummod	18	9	8	4
flat	19	11	11	3
advcl	20	21	21	23
conj	21	18	17	21
acl	22	23	22	18
csubj	23	22	24	22
appos	24	19	19	17

	F-score Rank	*Link Length*	*LISCA Gold*	*LISCA Parsed*
F-score Rank	1			
Link Length	0,42	1		
LISCA Gold	0,74	0,64	1	
LISCA Parsed	0,74	0,63	0,99	1

Figure 3
Ranking of the 24 DR types based on labelled F-score (*F-score Rank*), average *LISCA Score Gold* and *LISCA Score Parsed*, and average length of relations (*Link Length*) (the last three measures refer to the gold test set). Rankings Correlation tables report, for each language, Spearman correlation between the rankings (p <0.05 in all cases).

strong correlation values, ranging between 0.88 for Italian and 0.74 for English. Third, the DR list ranked by average dependency length shows the lowest correlation values for all languages with respect to the other metrics: note that the correlation turned out to be lower with the F-score than with the LISCA rankings. This result suggests that a structural property such as dependency length, despite being recognised as a well-known proxy of parsing difficulty, is not the only factor negatively affecting the parsing difficulty of dependency relations.

If we take a closer look at the rankings of DR types and their associated dependency length, interesting similarities and differences across languages can be observed. In most cases, the relative positions of the same DR types are very similar across rankings. Longer and difficult-to-parse DRs (i.e. with low F-score and LISCA scores) occur in the lowest part of all rankings. This is the case, for example, of three clausal relations: adverbial clause modifier (`ADVCL`) and clausal complement (`CCOMP`) for all languages, and clausal subject (`CSUBJ`) for Spanish and English. On the contrary, shorter and easier-to-parse DRs (i.e. with higher F-score and LISCA scores) are highly ranked in all languages. They mainly correspond to relations involving functional words, i.e. deter-

miner (DET), auxiliary (AUX) and case (CASE). This is not surprising, since it is a widely acknowledged fact that parsing systems have lower accuracy in the analysis of long dependency relations and, conversely, shorter relations are easier to parse (McDonald and Nivre 2007).

However, significant differences in terms of ranking positions also occur. Among them, it is worth reporting here difficult-to-parse DR types, i.e. characterised by low F-score and LISCA scores, but with lower average dependency length. This is the case, for all languages, of DRs involving nominal words that are core arguments of clausal predicates, i.e. nominal subjects (NSUBJ) and indirect objects (IOBJ), or relations linking multi-word expressions, e.g. FIXED. On the contrary, functional DRs involving words introducing subordinate clauses (MARK) or punctuation (PUNCT) are ranked higher in the F-score and LISCA orderings with respect to the ranking by average dependency length.

While DRs in the highest vs lowest part of the orderings could be considered as uniformly easy vs difficult to parse, the question which arises here is how the other DRs of the ranked list should be classified. Based on what we observed so far, we could imagine that they could be characterised by more variable contexts. However, the dependency type analysis presented so far doesn't allow to explore this intuition in depth and to understand which contextual features determine parsing difficulty. To address this issue, we need to move to a token-based (as opposed to type-based) analysis: Section 5 explores the properties contributing to make specific DR instances more or less difficult to be parsed.

5. Dependency-token Analysis

This second experiment is aimed at investigating, from both qualitative and quantitative perspectives, the linguistic properties of the context that make a specific DR instance more or less difficult-to-parse. In particular, it is aimed at addressing the open issues raised by RQ2 above. To investigate them, the DR ranking is used to:

a) single out difficult-to-parse DR instances from easy-to-parse ones;
b) identify the linguistic properties that contribute to make DR instances more or less difficult-to-parse and analyse their variation across languages.

5.1 Experimental Setting

In all experiments, each DR ranking was divided into 10 intervals of equal size, henceforth "bins", with the first bins containing DR instances presenting higher LISCA scores, and, conversely, the last bins with DR instances characterised by lower LISCA scores.

As a preliminary step, each DR instance in each bin has been checked for correctness. In the previous experiment, we showed that there is a high correlation between the LISCA-based ranking of DR types and parsing accuracy (measured in terms of F-score). Now we would like to narrow the investigation and focus on the relationship between the LISCA ranking of the DR instances and their correctness, i.e. whether they received a correct or wrong analysis when automatically parsed. In particular, we checked how parsing accuracy of DR instances ranked on the basis of the LISCA scores varies across different portions of the ranking.

For this purpose, we compared the ranked lists of DRs in the test set of each language obtained on the basis of the LISCA scores (both Parsed and Gold) and Dependency Length (used as a baseline). In the case of the DR ranking based on the *LISCA*

Score Parsed, the correctness has been established with respect to the gold version of the test set. Conversely, in the case of the ranking based on the *LISCA Score Gold*, the correctness has been assessed on the basis of whether a given gold DR instance shows the same analysis in the automatically parsed test set. For what concerns the DR ranking based on Link Length, DR correctness has been established with respect to the gold test set, as in the case of the *LISCA Score Gold*, i.e. on the basis of whether a given DR instance in the length based list is associated with the same analysis in the automatically parsed test set[5]. The results of such analysis are reported in Section 5.2 below.

In order to address a) and b) above, the distribution of both local and global characteristics of DR instances along the bins of the LISCA ranking was inspected, as reported in Section 5.3. In particular, the analysis focused on how two interrelated local features (i.e. the direction and linear distance between *h* and *d*) and a global feature (i.e. the number of dependents of *d*) vary across the bins, interact and contribute to make specific DR contexts difficult-to-parse. This analysis is carried out on the instances of nominal subjects occurring in the gold test sets of Italian and English languages. Specifically, the number of `NSUBJ` relation instances involved in the analysis is 500 and 1,863 for Italian and English respectively[6].

5.2 About the Correctness of Ranked Dependency Relation Instances

As mentioned above, the first step of our dependency-token analysis is aimed at investigating the relationship between the LISCA ranking of the DR instances and their correctness. Figure 4 summarises parsing accuracy results for each bin of the different DR rankings, including the length based one used as a baseline. To ensure comparability of the different DR distributions, relations are ordered for each language by increasing link length (see *Length* line) and by decreasing LISCA scores, both *LISCA Score Parsed* (see the *LISCA System* line) and *LISCA Score Gold* (see *LISCA Gold*). Intuitively, we expected correctly parsed instances to be placed higher in the LISCA rankings, for both the parsed and gold test sets, and, conversely, wrongly-parsed instances to be located in the lower part of the ranking or, anyways, in lower positions than their gold counterparts.

A common trend can be observed for all languages, namely the parsing accuracy decreases as we move progressively down the rankings: DR instances located in the top part of all rankings, when automatically parsed, are generally correct, whereas those occurring in the last bins are more difficult to parse, and thus more likely to be wrongly analysed.

A difference is reported for what concerns length based parsing accuracy lines. For all languages, they behave differently at the extremities of the ranking: in the top bins they show a lower accuracy with respect to the LISCA based rankings; on the other hand, in the bottom bins longer relations turned out to be more accurately parsed than relations characterised by low LISCA scores. This is in line with our previous findings, demonstrating that parsing difficulty is influenced by a variety of factors, and that dependency length – alone – is not sufficient to account for all aspects intervening in determining parsing complexity.

5 Note that the list of instances sharing the same length were internally ordered by appearance in the test set.

6 As we will discuss below, differences in the bin size across languages do not appear to affect our findings.

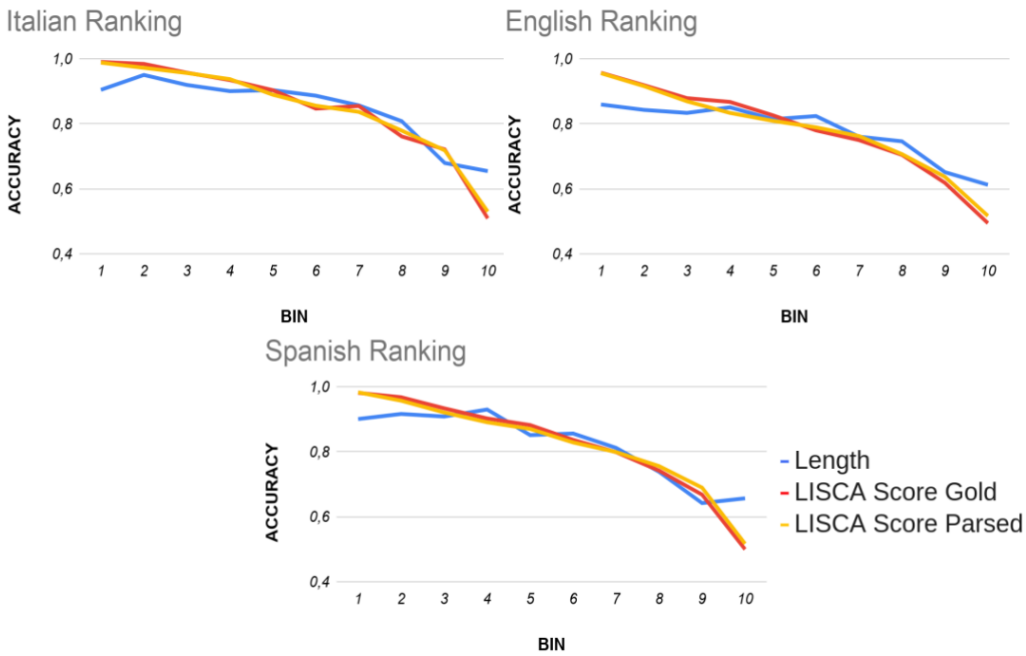


Figure 4
Parsing accuracy computed for each bin of the ranked list of DRs in the Italian, English and Spanish test sets. Relations are ordered by increasing link length, and by decreasing LISCA scores (i.e. *LISCA Score Parsed* and *LISCA Score Gold*).

Consistently with what reported for the first experiment (see correlation tables in Figure 3), the parsing accuracy resulting from the two LISCA rankings almost perfectly overlaps. Both LISCA rankings of DR instances reflect the gradual transition from easy- to difficult-to-parse DR contexts, and can thus be used to investigate the features contributing to make specific DR instances difficult to parse and how they vary across languages. The two different rankings can be used to highlight different contribution of LISCA to the topic of parsing difficulty. On the one hand, the ranking based on the *LISCA Score Parsed* can be used to identify difficult-to-parse and thus potentially wrong DR instances without using gold data. On the other hand, the ranking based on the *LISCA Score Gold* shows the LISCA potentiality to reliably identify syntactically complex dependency structures associated with a given DR type.

Starting from these premises, we are now able to take a step further in the analysis, i.e. to investigate which are the properties that characterise easy- vs difficult-to-parse contexts in which a given DR type occurs. We exemplify this type of analysis on the `NSUBJ` relation and two languages, i.e. Italian and English, belonging to two different typological genera. As a first step, the accuracy of `nsubj` instances across the ranked bins is considered: differently from the previous analysis, we focus here on the ranking based on the *LISCA Score Gold*, while we keep using *Link Length* ranking as a baseline. This decision naturally follows *i)* from the previously reported close similarity between the two LISCA rankings as well as *ii)* from our goal here, i.e. identifying the features which make specific contexts more or less difficult to parse, for which it is more sensible to exploit gold annotation data.

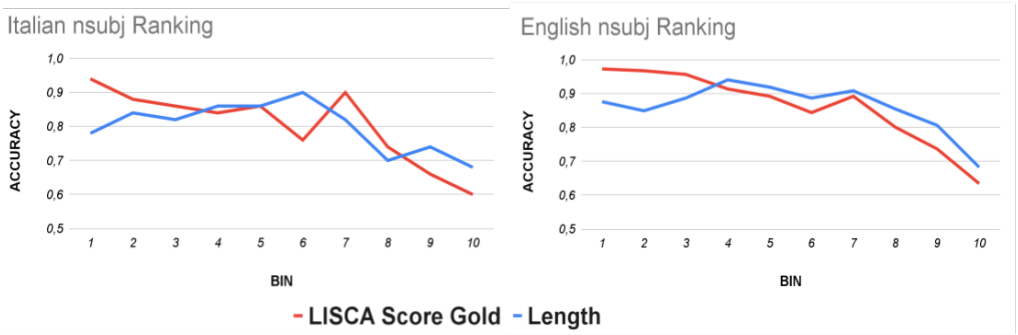


Figure 5
Parsing accuracy of Italian and English `nsubj` instances in different bins of the gold test set ordered by growing link length and decreasing LISCA score.

Figure 5 shows the accuracy of Italian and English nominal subjects across the rankings partitioned into ten bins of equal size. The picture emerging from the graphs is in line with what we previously observed for the ranking of all DR instances taken together: the first bins contain a higher number of correctly parsed instances with respect to the bottom ones. As in the previous case, the dependency length rankings show major differences at the extremes of the rankings. In order to better understand these results, we now consider the distribution of linguistic features widely known to contribute to parsing difficulty.

5.3 Singling-Out Difficult- and Easy-to-Parse Relations: Linguistic Properties Involved

Dependency Length and Direction. We start by monitoring the distribution of two local features which are recognised in the literature as contributing to linguistic and parsing complexity: namely, the linear distance from *d* to *h* (i.e, *dependency length*) and their relative ordering (henceforth, *dependency direction*).

While dependency length is widely known to determine the syntactic complexity of a sentence and thus to negatively affect its parsing accuracy (McDonald and Nivre 2007), dependency direction (namely, pre- or post-verbal subject position) reflects a language-specific property typically connected with "marked" or "unmarked" word orders (i.e. more or less prototypical, see (Haspelmath 2006)) which poses different challenges for what concerns parsing accuracy (Collins 2003; Gulordava and Merlo 2016). The graphs in Figure 6 report, for Italian and English, the distribution across the bins of `nsubj` instances on the basis of these two features; the last `TOT` column reports the same type of information for the whole DR set. The top graphs (a and c) focus on dependency length: although we can observe a common trend, Italian is characterised, on average, by longer relations with respect to English (2.9 is the average DR length for Italian, whereas it is 2.3 for English). In particular, Italian relations with length > 10 appear in the 6th bin and increase progressively to cover almost half (46%) of the cases in the last bin; conversely, in English they occur only in the in the last bin, covering around 15% of the cases. On the other hand, relations involving adjacent elements (i.e. with length equal to 1) show a similar distribution for the two languages only in the first bin: while in English short subject links decrease slowly but constantly from bin 1 to 10,

in Italian they decrease significantly starting from the 2nd bin, until they start covering less than 10% of cases from the 5th bin on. It should be noted that the overall number of these relations is much higher in English than in Italian: they cover 35% of English `nsubj` instances, while only 22% of the Italian ones (compare the TOT columns in the graphs).

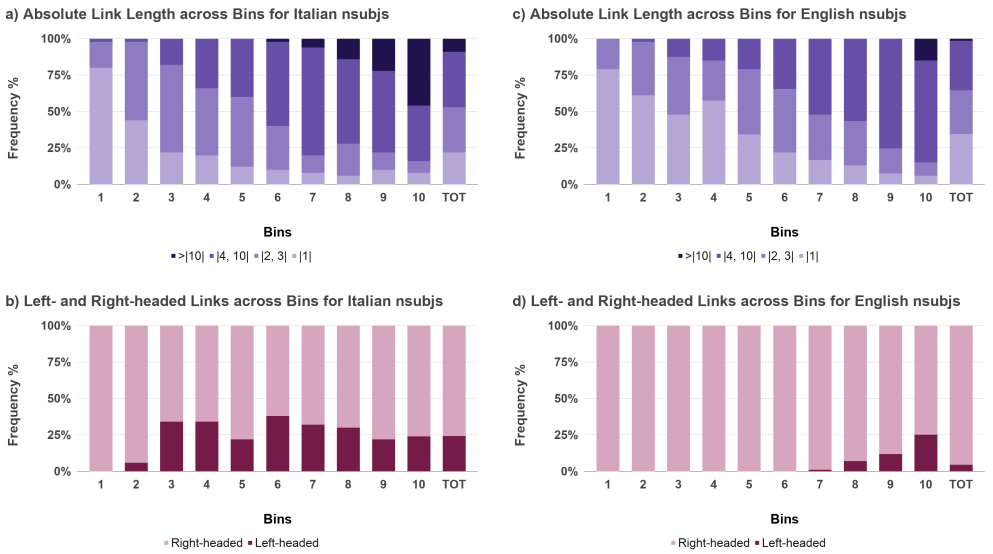


Figure 6
Relative frequency distribution across the LISCA bins of Italian and English `nsubj` instances with respect to absolute dependency length - graphs *a)* for Italian and *c)* for English - and link direction - graphs *b)* for Italian and *d)* for English.

Also the distribution of `nsubj` instances by dependency direction reported in graphs b) and d) of Figure 6 is quite different: left-headed subjects already appear in the second bin for Italian and, from the third one, start to cover a similar number of cases, ranging from 22% to more than 30%. In English they increasingly occur in the last 4 bins. In both cases, we are in front of a marked construction, which is however associated with a lower overall frequency in the language. It is interesting to note here that, although both languages are known to be SVO languages (i.e. preferring pre-verbal subjects), the marked option is differently distributed in Italian and English: it seems that Italian is generally more flexible, having a higher number of sentences showing the marked ordering option, while English data suggest that the language has a strong preference toward the unmarked subject-verb ordering.

About the Combined Effect of Linguistic Properties. The issue which remains open is how the two features interact and contribute to determine the position of a given DR instance along the ranking. Figure 7 shows the distribution of the two local features seen above across the bins and in the whole DR set (in the last `TOT` column).

The distribution of these features combined together across the LISCA bins provides a rich and articulated picture. Let us first look at the results from a monolingual perspective. For both languages, shorter right-headed (i.e. following the Subj-Verb order) links predictably concentrate in the first bins and, vice-versa, longer relations possibly following a "marked" order mainly occur in the bottom part of the ranking. For Italian,

very few instances of > 10-token long left-headed subjects can be found, all occurring in the last bin. For English, left-headed subjects concentrate in the last three bins and dependency length seems to be the main feature at play.

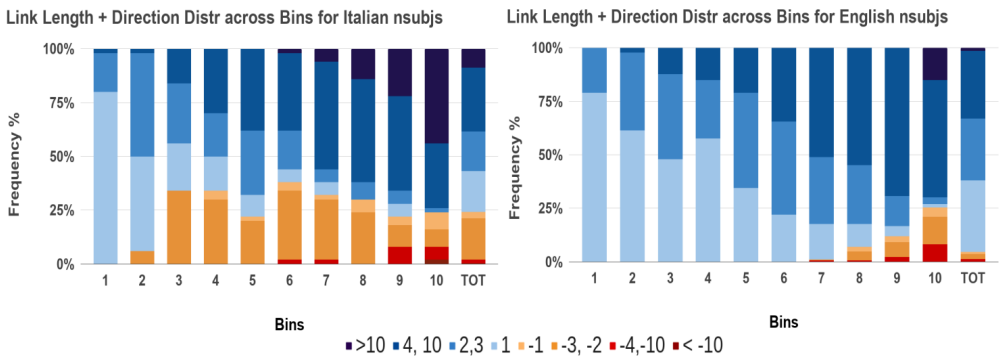


Figure 7
Relative frequency distribution across the LISCA bins of Italian and English `nsubj` instances with respect to dependency length and direction. Positive values identify right-headed relations, while negative values correspond to left-headed subjects.

On the other hand, the `TOT` column provides a flat distribution of the characteristics over the entire DR set from which it is not easy to assess what makes a subject instance difficult or easy to parse: although we can observe a different frequency of right- vs left-headed `nsubj` instances of a given length, we can't always tell how these local properties affect the parsing difficulty.

Consider now the distributions in Figure 7 from a multilingual perspective: it can be noticed that pre-verbal subjects reflect the canonical unmarked order in both Italian and English and are thus more frequent in the higher part of the ranking for both languages. However, it is interesting to report that for Italian, which is characterised by a higher word-order flexibility with respect to English, left-headed `nsubj` instances (represented as negative links in the graphs) already appear in the first half of the ranking. This is not the case for English, where left-headed `nsubj` instances concentrate in the last bins only.

The evidence emerging from Figure 7 is not surprising if we consider the typological properties of the two languages. However, it provides interesting information for what concerns the features contributing to make a specific DR instance complex to be automatically analysed. The fact of finding cases of short links in the last bins, and some long relation instances appearing earlier in the ranking suggests that there are different factors at play which interact in determining the parsing complexity of a relation instance. Interestingly, the interaction of the two properties turned out to vary across Italian and English.

Similar observations hold also in the case of global structural features, such as the number of dependent elements (i.e. "children" nodes) of d, as shown in Figure 8. By inspecting the overall distribution reported in the `TOT` columns we can observe that the majority of `nsubj` instances are represented by leaf nodes in the English case, whereas they are less than one third in Italian. This situation follows from a syntactic property distinguishing Italian from English, namely the fact that subjects can be omitted: this results in a much higher frequency of pronominal subjects in English, which always explicitly expresses the subject, with respect to Italian where it can be omitted. If

frequency can thus be used in English to identify easier to parse `nsubj` instances, this is not the case for Italian where it is rather the distribution of `nsubj` instances represented by leaf nodes across the LISCA bins to provide useful evidence concerning their parsing accuracy, which is expected to be higher with respect to `nsubj` instances with children nodes.

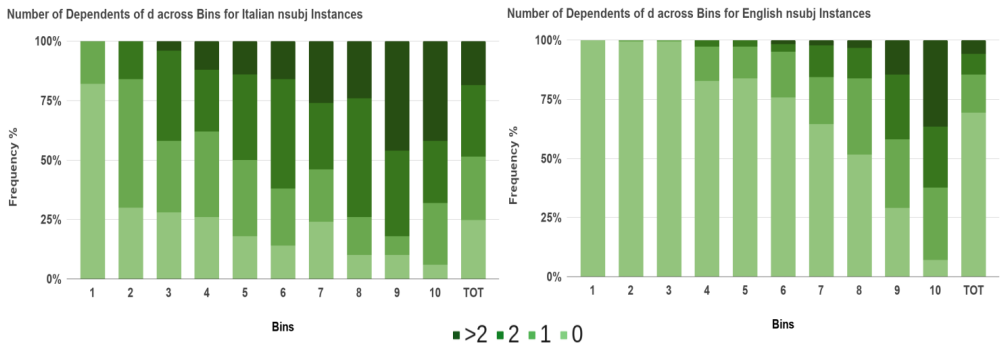


Figure 8
Relative frequency distribution across the LISCA bins of Italian and English `nsubj` instances with respect to their number of dependants.

To better appreciate how the wider context of a DR instance is captured by the LISCA score, consider e.g. the following sentences containing an English subject (in italic) and the relation with its syntactic head (underlined): a) '*I* had to go to the BBC for this report', and b) 'An orphaned, two-month old African *elephant* named Olly received an extremely uplifting Christmas present this year [...]. Sentence a) presents a typical case of English subject: a pronoun immediately preceding its syntactic head. Such prototypical unmarked subject cases can be found in the higher part of the ranking; more precisely, this example was taken from the first bin. On the other hand, the `nsubj` in b) represents a tree-token long DR which has in turn six children and which is found in the 9th bin of the ranking. Interestingly, the different levels of difficulty are confirmed by the automatic parsing results (with UDpipe) of these sentences, displayed in Fig. 9: because of the different contexts where they occur, the two `nsubj` instances are differently parsed, with the former which is correctly parsed, and the latter resulting in an error.

To conclude, the results outlined above complement and expand what already observed in the first experiment. First, they show that the LISCA score can reliably be used to identify difficult-to-parse constructions and that this could be done without relying on gold data. Second, the variety of local and global properties captured by the features taken into account make the LISCA score a context-sensitive measure to predict parsing difficulty, which also permits to identify and weight the contribution of individual features in determining the parsing difficulty of specific constructions.

6. Discussion and Conclusion

This paper represents a contribution to the studies devoted to define a method to quantify dependency parsing performance. The dependency-type and dependency-token analyses proposed in the previous sections provide preliminary but promising

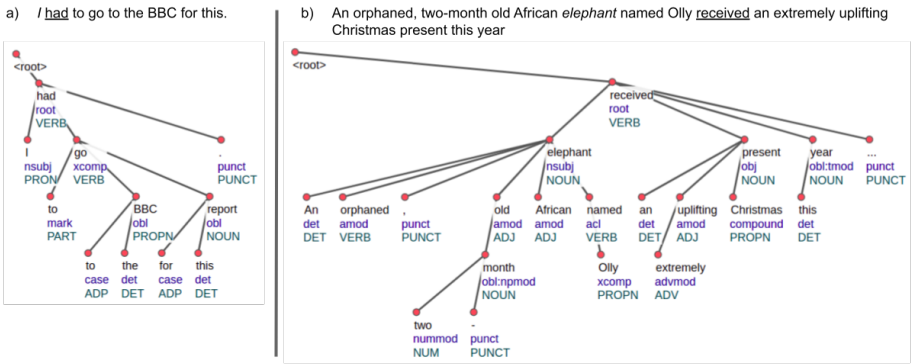


Figure 9
Dependency trees representing the linguistic analysis of sentences a) and b).

evidence to answer the complementary and intertwined research questions we started with.

For what concerns the first question (RQ1), the results of the experiments reported in Section 4 demonstrated that the LISCA-based ranking of dependency relations is highly correlated with the ranking obtained using the most common measure of parsing reliability, i.e. the labelled F-score, but with a main difference: the LISCA-based ranking does not need gold data. This represents a first innovative feature of our approach, which is able to assess the degree of parsing complexity by solely relying on local and global characteristics of the linguistic context where DRs occur. The wide variety of properties taken into account by LISCA thus turned out to be a reliable predictor of parsing difficulty by DR type.

The second research question (RQ2), investigated with the experiments reported in Section 5, was aimed at exploring whether and to what extent the variability of linguistic contexts in which the same DR type occurs contributes to the positioning of a specific DR instance along the ranking. We showed that, for a given DR type, LISCA is able *i)* to single out easy vs difficult-to-parse contexts by relying on both local and global properties associated with the single DR instance, and *ii)* to identify and weight the features making it more or less difficult-to-parse together with their interaction. In particular, we demonstrated how two well-known properties, i.e. the length and the direction of a dependency relation, interact with each other in determining the parsing difficulty of `nsubj` instances. This represents another important novelty of our approach, which anchors complexity features to specific DR types and also investigates their interaction. Reported results demonstrate that typology and interaction of complexity features can significantly vary across languages.

As pointed out in the introductory section, quantifying the parser performance with respect to DR types and weighting the degree of parsing difficulty on the basis of the contextual properties represent different and complementary lines of research, which are typically carried out separately by resorting to different methods and techniques. The LISCA-based approach proposed here turned out to be suited to tackle both perspectives of analysis by relying on the same methodology. We also showed that this methodology can be effectively applied to different languages, representative of different typological genera. Current developments are concerned with the application of this methodology to low-resourced languages, for which large amounts of data are

difficult to acquire and which thus need alternative solutions for obtaining the LISCA model without relying on large corpora.

On the basis of results achieved so far, we believe that our methodology could be usefully exploited in different application scenarios. First, it could represent the starting point to devise a new context-sensitive evaluation metric able to identify complex constructions, that might result in parsing errors, even without resorting to gold data. Second, the method could also be used in a resource building scenario, to build challenge sets based on the selection of sentences that contain difficult-to-parse constructions. The automatic extraction of challenge sets makes the proposed approach a scalable and practical solution for evaluating dependency parsing performance on the long-tail of complex syntactic phenomena. In particular, it represents an alternative solution to the traditional approaches that rely on hand-crafted rules formalising *a priori* knowledge or selecting dependency relation types without further refinements (Naseem, Barzilay, and Globerson 2012; Täckström, McDonald, and Nivre 2013; Scholivet et al. 2019). Differently from these approaches, the method can be used to select sentences that contain dependencies sharing linguistic constructions characterised by similar degree of parsing difficulty. This would allow to automatically create linguistically motivated test suites to be exploited as benchmarks for the evaluation of parsing systems, or, alternatively, as training sets including sentences that share an homogeneous complexity degree.

References

Alonso, Héctor Martínez and Daniel Zeman. 2016. Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020. Quantitative linguistic investigations across universal dependencies treebanks. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it)*, Bologna (online), Italy, March.

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 201–210, Prague, Czech Republic, January.

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4540–4549, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2019. Inferring quantitative typological trends from multilingual treebanks. A case study. *Lingue e Linguaggio*, XVIII(2):209–242.

Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Collins, Michael. 2003. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

Dell'Orletta, Felice, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistically-driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 17(2):125–136.

Droganova, Kira, Daniel Zeman, Jenna Kanerva, and Filip Ginter. 2018. Parse me if you can: Artificial treebanks for parsing experiments on elliptical constructions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Frazier, Lyn. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen, and A.M. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK.

Gulordava, Kristina and Paola Merlo. 2015. Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, Beijing, China, July. Association for Computational Linguistics.

Gulordava, Kristina and Paola Merlo. 2016. Multi-lingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of linguistics*, 42(1):25–70.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Number 73. Cambridge University Press.

Jakobson, Roman. 1973. *Essais de linguistique générale t. 2: rapports internes et externes du langage*. Les éditions de Minuit.

McDonald, Ryan and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131, Prague, Czech Republic, June. Association for Computational Linguistics.

McDonald, Ryan and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Merlo, Paola. 2015. Evaluation of two-level dependency representations of argument structure in long-distance dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 221–230, Uppsala, Sweden, August. Uppsala University, Uppsala, Sweden.

Naseem, Tahira, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.

Nivre, Joakim. 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.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Nivre, Joakim and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden, May. Association for Computational Linguistics.

Rimell, Laura, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821, Singapore, August. Association for Computational Linguistics.

Scholivet, Manon, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3919–3930, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Straka, Milan, Jan Hajič, and Jana Straková. 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'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia, June. Association for Computational Linguistics.

Tusa, Erica, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 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.

Yngve, Victor H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.