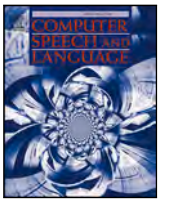
Computer Speech and Language 71 (2022) 101261

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



BERT syntactic transfer: A computational experiment on Italian, French and English languages

Raffaele Guarasci [a,*], Stefano Silvestri [a,*], Giuseppe De Pietro [a], Hamido Fujita [b], Massimo Esposito [a]

[a] Institute for High Performance Computing and Networking of National Research Council of Italy (ICAR-CNR), via Pietro Castellino 111, 80131, Naples, Italy

[b] Iwate Prefecture University, Takizawa, Iwate, Japan

ARTICLE INFO

Keywords:
Cross language
Dependency Parse Tree
Language models
Multilingual BERT
Transfer learning
Syntactic phenomena

ABSTRACT

This paper investigates the ability of multilingual BERT (mBERT) language model to transfer syntactic knowledge cross-lingually, verifying if and to which extent syntactic dependency relationships learnt in a language are maintained in other languages. In detail, the main contributions of this paper are: (i) an analysis of the cross-lingual syntactic transfer capability of mBERT model; (ii) a detailed comparison of cross-language syntactic transfer among languages belonging to different branches of the Indo-European languages, namely English, Italian and French, which present very different syntactic constructions; (iii) a study on the transferability of a syntactic phenomenon peculiar of Italian language, namely the pronoun dropping (pro-drop), also known as omissibility of the subject. To this end, a structural probe devoted to reconstruct the dependency parse tree of a sentence has been exploited, representing the input sentences with the contextual embeddings from mBERT layers. The results of the experimental assessment have shown a transfer of syntactic knowledge of the mBERT model among these languages. Moreover, the behaviour of the probe in the transition from *pro-drop* to *non-pro-drop* languages and vice versa has proven to be more effective in case of languages sharing a common linguistic matrix. The possibility of transferring syntactical knowledge, especially in the case of specific phenomena, meets both a theoretical need and can have important practical implications in syntactic tasks, such as dependency parsing.

1. Introduction

Characterising mechanism through which different aspects of linguistic knowledge can be transferred among different languages has always fascinated scholars from different fields of research. Many studies in past years have focused on differences across languages considering semantic aspects. Recent studies (Majid et al., 2015; Thompson et al., 2018) have analysed closeness of semantic spaces across different languages, showing that two languages are more semantically aligned the closer they are phylogenetically. Further evidence of the possibility of shared semantic spaces comes from computational semantics (Hauer and Kondrak, 2020; Camacho-Collados et al., 2015).

But the area that can most benefit from recent cross-lingual line of research in Natural Language Processing (NLP) is Syntax (Linzen and Baroni, 2021; Dhar and Bisazza, 2020). From the very beginning, one of the main aims of syntactic linguistic

* Corresponding authors.
E-mail addresses: raffaele.guarasci@icar.cnr.it (R. Guarasci), stefano.silvestri@icar.cnr.it (S. Silvestri), giuseppe.depietro@icar.cnr.it (G. De Pietro), issam@iwate-pu.ac.jp (H. Fujita), massimo.esposito@icar.cnr.it (M. Esposito).

https://doi.org/10.1016/j.csl.2021.101261
Received 5 January 2021; Received in revised form 29 April 2021; Accepted 2 June 2021
Available online 1 July 2021
0885-2308/© 2021 Elsevier Ltd. All rights reserved.

theory has been to identify general principles that recur in every language, defined as linguistic universals (Comrie, 1989; Gass, 1984; Newmeyer, 2008; Croft, 2009).

Syntactic features differ from one language to another, but changes are not arbitrary. Linguistic theory has already assumed that generalisations are allowed up to a certain point: some constructions will be different, others will be the same (Chomsky, 1981). This insight has been confirmed by recent studies in the field of Neurolinguistics (Declerck et al., 2020; Hartsuiker et al., 2016, 2004) and neural models of language (Conneau and Lample, 2019; Chi et al., 2020; Conneau et al., 2020b).

Starting from the assumption that abstract syntax trees (and its computational implementation called Dependency Parse Tree) can represent every kind of syntactic description in every language (McCoy et al., 2020; Kondratyuk and Straka, 2019; Kolachina and Ranta, 2019), in NLP field shared syntactic descriptions for multiple languages have been proposed (Ranta et al., 2009; Nivre et al., 2016, 2020b), but many problems still remain unresolved.

In recent years NLP has undergone profound changes. High-performance Deep Learning architectures have involved every task of NLP, ranging from sentiment analysis (Li et al., 2020) to text classification (Du et al., 2020) or anaphora and coreference resolution (Sukthanker et al., 2020). Newborn Deep Learning Language Models (NLMs) such as ELMo (Peters et al., 2018) or Transformer (Vaswani et al., 2017) based architectures, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), have not only improved state-of-the-art performance in several NLP tasks, but they have also shown that they can encode linguistic knowledge.

In particular, the authors of Tenney et al. (2019a) have studied where linguistic information related to syntactic and semantic structure is captured within the layers of the BERT network, by exploiting the edge probe approach (Tenney et al., 2019b). The probe aims at measuring how well information about linguistic structure can be extracted from a pretrained encoder by decomposing structured-prediction tasks into a common format, where a probing classifier receives spans and must predict a label such as a constituent or relation type. Their results provided evidences corroborating that BERT model can represent the types of syntactic and semantic abstractions in an interpretable and localisable way, and that the regions responsible for each step appear this sequence: Part of Speech (PoS) tagging, parsing, Named Entity Recognition (NER), semantic roles and finally coreference.

Moreover, the authors of Jawahar et al. (2019) applied the probing approach proposed by Conneau et al. (2018) to show that BERT representation embeds phrase-level information in its lower layers and hierarchy of linguistic information in its middle layers, with surface features at the bottom, syntactic features in the middle and semantic features at the top. In addition, they found that BERT requires deeper layers when long-distance dependency information such as subject–verb agreement is required.

Due to the great success of these models in monolingual NLP tasks, recent studies have also opened up to work on multiple languages. The recent introduction of multilingual NLMs at scale, such as mBERT (Devlin et al., 2019) or XLM (Conneau et al., 2020a), has allowed to perform a wide range of cross-lingual natural language tasks, such as Named Entity Recognition, Part of Speech Tagging, Neural Machine Translation, Text Classification (Pires et al., 2019; Silvestri et al., 2020; Conneau et al., 2020a; Wu and Dredze, 2019; Esuli et al., 2020; Hajmohammadi et al., 2015; Catelli et al., 2020; Siddhant et al., 2020; Pamungkas et al., 2020), also with few-shot or zero-shot learning approaches (Hayashi and Fujita, 2020). More interesting, it has also been shown that the ability of mBERT to generalise across different languages does not simply rely on vocabulary memorisation, but it is able to learn a deeper multilingual representation (Pires et al., 2019), in particular when the different languages show similar linguistic structures and typological features. But an open issue not widely investigated until now regards what happens when the languages show completely different features of constructs.

1.1. Research objectives

This paper has two main objectives. Firstly, it is aimed at testing the capability of a multilingual NLM to transfer syntactic features across different languages. Secondly, it investigates if a specific syntactic phenomenon peculiar of a single language can be correctly cross-lingually embedded into mBERT layers. To this end, the mBERT model has been experimented to assess if and how it is able to transfer syntactic dependency relationships across three different languages, namely Italian, English and French.

Otherwise, the choice of languages under consideration here has precise reasons deriving from Linguistic Typology. Although belonging to the Indo-European language family, they present some noteworthy typological variations and a series of differences, both at the level of the word order and on specific phenomena. English is a Germanic language that has static syntactic structures, with a mandatory expressed subject, while Italian and French are part of the Italic languages (in particular the Romance languages sub-group). However, French is one of the few Romance languages in which subject must be expressed. By contrast, Italian is characterised by a high degree of word order freedom (Futrell et al., 2015). The syntactic structure is very variable and the pronoun subject *de facto* never made explicit, as it occurs in other Romance languages like Spanish (Lahousse and Lamiroy, 2012).

This specific phenomenon, named pronoun-dropping (i.e. the omissibility of the subject pronoun) has been also the object of the proposed analysis. Unlike other works proposed so far, evaluation process is not limited to metrics proposed by Hewitt and Manning (2019), but it also includes a qualitative analysis, involving native-speakers perspective. Indeed, the transition from a language such as English or French with the mandatory expressed subject to a language where in most cases it can be omitted, like the Italian, is challenging for the way syntactic relations are reconsidered using a NLM.

Although from a strictly linguistic point of view it certainly makes more sense to compare only closely related languages (Søgaard et al., 2018; Vulić et al., 2019), limiting the analysis to Romance languages French and Italian would reduce the possible benefits of the cross-lingual approach. One of the main purposes of modern cross-lingual approaches is to exploit the great availability of resources in languages such as English in order to open up research perspectives and boost performance of NLP systems in less-resources languages, such Italian (Wu and Dredze, 2019; Conneau et al., 2020b; Cruz et al., 2018).

1.2. Contributions

This paper provides the following main contributions: (i) an assessment of the multilingual BERT cross-lingual syntactic transfer capability; (ii) a comparison among English–French, English–Italian, Italian–English, Italian–French, French–English and French–Italian cross-language syntactic transfer leveraging mBERT model; (iii) a study on the transferability of *pro-drop* (omissibility of the subject) syntactic phenomenon.

In detail, a structural probe (Hewitt and Manning, 2019) has been used to approximate syntactic dependencies in the form of unlabelled DPTs. The probe has been first trained on a language and, then, tested on the two other ones, by considering all the possible combinations. A quantitative analysis has been performed to determine the extent to which approximations of syntactic relationships embedded in the model can be transferable cross-lingually and the best layers of the model's internal representation embedding this transferred linguistic knowledge.

It is worth noting that, currently, there are only few works involving multiple languages (Wu and Dredze, 2019; Karthikeyan et al., 2020), but none of them takes into account these three languages in particular. The authors of Jawahar et al. (2019) have proposed to investigate mBERT syntactic tree representations focusing on isolated arguments (e.g. subject vs object) or specific linguistic categories (e.g. Determiners, Adjectives, Negatives).

A recent work (Chi et al., 2020) has studied how the syntactic relations of the model can be approximated in languages other than English. The work does not focus on specific syntactic phenomena but on the transfer from one language to another of Universal Dependency relations. It is part of the broader scenario of universal grammar related to the hypothesis of shared grammatical relations between all languages. However, although their analysis brings together several very different languages, Italian is not included. Moreover, it based the evaluation only on Universal Dependency (Nivre et al., 2020a) relations. As far as known, there are no specific studies dealing with the Italian, even though there are now several pretrained BERT models for this language.

1.3. Outline

The rest of this paper is structured as follow. The Section 2 provides an overview of the recent related works. The next Section 3 is devoted to the description of the research methodology, as well as the details of the structural probe, the Neural Language Model and the syntactic task and phenomena investigated. The Section 4 describes the performed experimental assessment, also providing the information related to the datasets and the adopted metrics. In Section 5 the obtained results are presented and discussed. Finally, Section 6 summarises the paper and draw out the final conclusions.

2. Related works

Theoretical Linguistics have defined syntax as an abstract mechanism in which combinatorial operations bound by precise rules regulate the use and relations between words (Chomsky, 1995). These rules can rely on lexical-phonological associations (Thierry and Wu, 2007) or syntactic relations (Loebell and Bock, 2003; Shin and Christianson, 2009; Gries and Kootstra, 2017).

The possibility of a shared syntax (or syntactic representation) is a well researched topic in Theoretical and Computational Linguistics. From a Theoretical point of view, this line of research is rooted in the hypothesis of Chomsky's Universal Grammar (Chomsky, 1957). The topic has also been widely researched in Neurolinguistics (Hartsuiker et al., 2004), suggesting evidences for a shared syntactic representation useable across different languages. Few empirical studies have dealt with the differences and similarities between the different grammars, classifying the differences as coarse-grained on the basis of theoretical considerations (Dorr et al., 1994).

Concerning NLP, scholars have been interested in cross-lingual syntactic studies since the dawn of the field. In particular, machine-translation studies have assumed that the syntactic structure of a sentence can be predicted using the syntactic structure of its translation (Kozhevnikov and Titov, 2013; Rasooli and Collins, 2017).

With the rise of language models based on neural networks and their success on monolingual tasks (Tenney et al., 2019a; Jawahar et al., 2019), scholars have also begun to question multilingual possibilities. In Wu et al. (2020) and Pires et al. (2019) the authors have tested the cross-lingual potential of the Multilingual version of BERT (Devlin et al., 2019), achieving surprisingly interesting results using mBERT to make generalisations between different languages. The model has been tested also on specific NLP tasks (i.e. Named Entity Recognition and Part of Speech Tagging). In Conneau et al. (2020a) it has been shown that the XLM NLM outperforms mBERT in cross-lingual Natural Language Inference, Question Answering and other NLP tasks.

The authors of Rönnqvist et al. (2019) have explored mBERT ability in language generation, finding that English and German models perform well at generation, whereas the multilingual model is lacking for Nordic languages. In Karthikeyan et al. (2020), a study on different NLP tasks has been presented, taking into account three typologically different languages. Results have shown that the crucial role in cross-lingual performance is done by lexical similarity between languages. In Pires et al. (2019) it has been demonstrated that mBERT is able to generalise cross-lingually without being explicitly trained for it, therefore a specific multilingual training can significantly increase performances (Conneau and Lample, 2019).

With reference to more specific experiments on the approximation of syntactic phenomena using neural networks, it is worth remembering some recent works. The possibility of a neural model (Transformer-based like BERT) to learn some form of syntactic knowledge exploiting structures and dependencies is a relatively recent interest in the field of NLP (Warstadt et al., 2019). The authors of Clark et al. (2019) have probed the attention heads of the BERT architecture for linguistic phenomena, treating each attention head as a simple no-training-required classifier that, given a word as input, outputs the most-attended-to other word. In

this way, they have evaluated the ability of the heads to classify various syntactic relations, finding that particular heads correspond remarkably well to particular relations, such as direct objects of verbs, determiners of nouns, objects of prepositions, and objects of possessive pronouns, as well as to coreference resolution. As further confirmation of this increased interest in syntax, Tenney et al. (2019a) has proved that BERT encodes syntax more than semantics.

The authors of Jawahar et al. (2019) have performed several experiments to discover which elements of English language structure are learned by BERT model. Results have highlighted that BERT different layers capture diverse levels of language complexity. Lower layers encode phrase-level information, middle layers deal with syntactic features and higher ones focus on semantic features. Deeper layers are required in order to perform task related to long-distance dependency information, such as subject–verb agreement.

In Hewitt and Manning (2019) a structural probing model has been trained with the purpose of showing that learned spaces of language models such as BERT and ELMo are better for reconstructing dependency trees than baselines. In particular, the hidden representations of each token into an inner-product space correspond to the distance of the syntax tree. The same probing approach has been exploited in Chi et al. (2020) to examine the extent to which mBERT learns a cross-lingual representation of syntactic structure. Their experimental assessment in 11 different languages has provided the evidence that mBERT shares at syntactic level some portions of its representation space between languages.

More recently, some studies have investigated these phenomena from a typological point of view. The authors of Bjerva and Augenstein (2021) have formulated the hypothesis that a mBERT model is able to embed typological information from the input data. They have verified their hypothesis by blinding mBERT model to typological information (syntactic, morphological and phonological) by using gradient reversal technique (Ganin and Lempitsky, 2015). Then, they have evaluated PoS (Part of Speech) tagging, NER (Named Entity Recognition), XNLI (cross-lingual Natural language Inference), and PAWS-X (paraphrase identification) tasks on 40 different languages. The obtained results have proved that preventing this model from exploiting typology severely reduced performance, while exposing the model to it resulted in increased performances, especially in case of syntactic and morphological information.

In Ravishankar et al. (2021) decoding experiments for mBERT across 18 languages have been presented. The purpose of these experiments is testing if dependency syntax is reflected in attention patterns. In particular, the main aim is to confirm that the attention patterns of BERT-based models can capture structural features across typologically diverse languages, as observed in English. To this end, dependency tree decoding algorithm (Raganato and Tiedemann, 2018) over every layer and head combination of mBERT model in 18 different languages have been run, demonstrating that the mBERT is able to decode dependency trees from attention patterns more accurately than an adjacent-linking baseline, implying that some structure was indeed being tracked by the mechanism.

It is worth noting that the vast majority of these works are for English language, or do not focus on specific syntactic phenomena. A study carried out on Spanish relative clauses has been presented in Davis and van Schijndel (2020), showing that non-linguistic prejudices in Recurrent Neural Network Language Models overlap the syntactical structure in English, but not in Spanish. Although works in recent years have examined many different languages, currently, no syntax transfer experiment has been conducted simultaneously involving the three languages taken into account in this work. Particularly with regard to the Italian language, there is no work examining the approximation of syntactical knowledge based on a language neural model.

3. Methodology

This study focuses on cross-lingual syntax transfer using three different languages, namely English, and French. The principal aim is to validate the extent to which a NLM trained on a language can learn syntactic information, in particular dependency relations (expressed by DPTs), in its contextual word representations and transfer it cross-lingually to another language, adopting the structural probe proposed by Hewitt and Manning (2019) and considering mBERT as language model. This syntax-transfer based on mBERT is consistent with hypotheses like Universal Grammar and shared syntax.

For the sake of clearness, in Fig. 1 an overview of the methodological aspects covered by this paper is given. In detail, the three datasets indicated on the left represent language-specific, syntactically annotated datasets to be used for training the language model, whereas the three datasets indicated on the upper side are language-specific parallel datasets to be used for testing the language model. The output is given by the approximation of syntactic dependency relationships (DPT). The central block includes the building elements used to configure the experiments: (i) structural probe, (ii) the language model and (iii) syntactic task and phenomena to be analysed as detailed in the yellow balloons and covered hereinafter.

3.1. Structural probe

Probes are supervised models designed to test hypotheses, given a specific phenomenon. They extract a linguistic structure from the output representation learned from a model, providing an evidence of a phenomenon. The probes also provide, at the same time, a way to extract the phenomenon of interest from the model. In Hewitt and Liang (2019) the authors focused on how to design and interpret probes. The structural probe here adopted has been previously presented by Hewitt and Manning (2019), demonstrating that monolingual BERT model encodes in its layers the syntactical tree structures of the sentences.

The structural probe takes, as input, the sequence of contextual embeddings corresponding to the n words $w^m_{1:n}$ of the sentence m, producing, as output, a sequence of vector representations $\mathbf{h}^m_{1:n}$. A tree structure is embedded if this transformed space has the property that squared L2 distance between two word vectors corresponds to the number of edges between the words in the parse

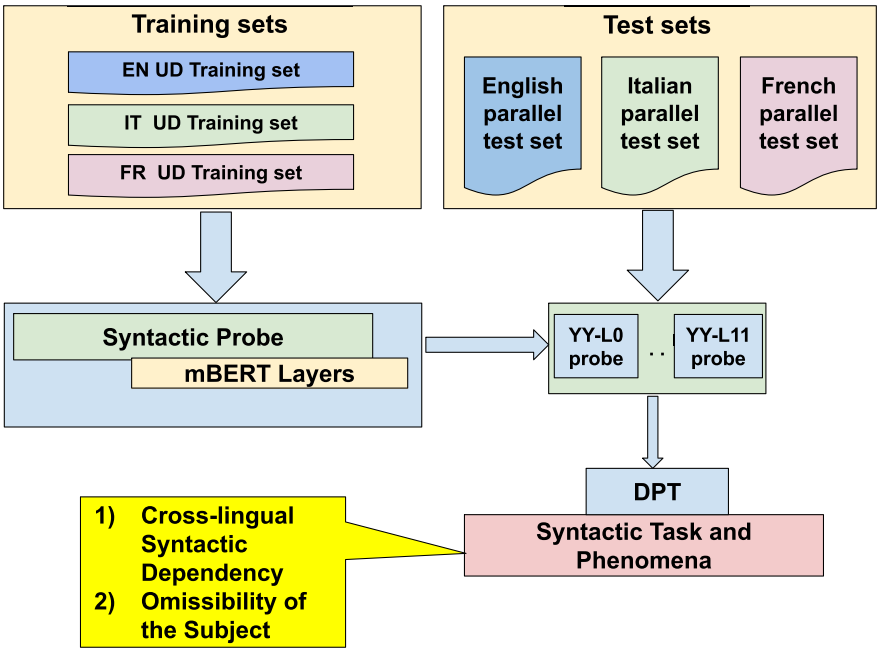


Fig. 1. Methodology overview. The figure highlights an overview of the methodological aspects covered by this paper. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tree. Thus, the probe is defined through an inner product on the original space under which squared distances and norms encode syntax trees. Exploiting the dot product properties, it is possible to define a family of inner products $\mathbf{h}^T A\mathbf{h}$, parameterised by any symmetric, positive semi-definite matrix A.

A linear transformation B can be defined as $A = B^T B$ and the inner product can be expressed as $(B\mathbf{h})^T(B\mathbf{h})$, which is also the norm of $\mathbf{h}$ transformed by B. Then, it is possible to define for each sentence a family of squared distances:

$$d_B(\mathbf{h}_i^m, \mathbf{h}_j^m)^2 = (B(\mathbf{h}_i^m - \mathbf{h}_j^m))^T(B(\mathbf{h}_i^m - \mathbf{h}_j^m)) \tag{1}$$

where i, j are the indexes of the word in the sentence m.

The trainable parameters of the probe are the coefficients of the matrix B, which are trained to predict the tree distance between all words for each sentence of the training set by solving the following optimisation problem:

$$\min_B \sum_m \frac{1}{|s^m|^2} \sum_{i,j} |d_{T^M}(w_i^m, w_j^m) - d_B(\mathbf{h}_i^m, \mathbf{h}_j^m)^2| \tag{2}$$

where $|s^m|^2$ is the square of the length of the sentence mth.

This probe defines a valid distance metric, which is non-negative and symmetric and, furthermore, it tests that there exists an inner product on the representation space whose squared distance encodes syntax tree distance. In this way, it allows the model to encode not only which word is related to which other word, but also each word's proximity to every other word in the syntax tree and produces a parse-tree-like representation in output (for more details about the probe implementation, as well as parameters and hyperparameters settings, please refer to Hewitt and Manning, 2019).

3.2. Neural language model

The language model tested through the probe is mBERT,[1] a Transformer-based (Vaswani et al., 2017) NLM, which exploits the BERT (Devlin et al., 2019) architecture, and it is pretrained on a multilingual corpus formed by raw Wikipedia text from 104 different languages. In detail, the mBERT model uses the classical *BERT-Base* architecture, formed by 12 encoder-only Transformer layers, with 768 hidden dimensional states and 12 attention heads, counting approximately $110M$ parameters.

Notice that its aim is not focused on cross-lingual tasks. The training corpus does not use any marker denoting the type of input language and does not have parallel or aligned languages with mechanisms to support the translation-equivalent pairs to have similar representations. The mBERT original purpose is simply to be used as a universal language model and as a tool for encoding sentences in more languages.

[1] https://github.com/google-research/bert/blob/master/multilingual.md.

Table 1
Examples of possible word order in Italian. Sentences are a modified version of those originally presented in Bates et al. (1982).

VSO	Allora mangio anche io la pizza	*Well then, am eating also I pizza*
VOS	Ha consigliato la pizza qui Franco	*Recommended the pizza here Franco*
OVS	No, la pizza l'ha consigliata Franco	*No, the pizza recommended Franco*
SOV	Allora, io gli spaghetti prendo	*In this case, I the spaghetti am having*
OSV	La pizza la prendo sempre qui	*Pizza (I) order always here*
SVO	*È* stata incoronata qui	*(She) was crowned here*

The NLM is used to create contextual embeddings representing the sentences of the dataset, exploiting each one of the 12 layers of the model. In the experimental assessment, the mBERT Multilingual Cased model[2] is used.

3.3. Syntactic task and phenomena

The syntactic task here chosen to be investigated for assessing the capability of syntactic transfer learning of mBERT consists in the approximation of dependency relationships in sentences. This task is extremely relevant for the three languages here considered, since they present several syntactic differences. From a strictly linguistic point of view, according to the traditional classification (Blake, 1988), all three languages fall into the typology of the Subject–Verb–Object (SVO) family. However, even if the dominant word order (Whaley, 1996; Dryer, 2005) of Italian, English and French is the same, comparative studies have quantitatively analysed the extent of word order freedom across languages (Futrell et al., 2015; Liu, 2010).

Italian is a morphologically rich language (Tsarfaty et al., 2010) characterised by a high verbal inflection. Several studies have demonstrated that inflectional properties and variations of a language are strictly related to its syntactic properties (Liu and Xu, 2012), leading to a huge number of possible word forms, relatively free constituent order and pro-drop phenomenon (Alicante et al., 2012). As pointed out by the fundamental work of Bates et al. (1982), Italian allows all possible orders of subject, verb and object and this is an aspect that differentiates it fundamentally from English. The grammaticality of all these possible orders is often due to the deletion of the subject, which occurs in 70% of cases (Bates, 1974). In Italian – as for other Romance languages such as Spanish – alternative orders and free inversion can be considered as combinations of sentence fragments with deleted elements (Burzio, 1986; Rizzi, 1982).

Table 1 shows examples of every possible alternative word order in Italian sentences. The first column indicates the word order of the sentence, notice that in translated sentences in the last column the order of the original Italian sentences has voluntarily maintained, even if ungrammatical in English. Last two sentences contain pro-drop constructions in which the subject pronoun is unnecessary because it is grammatically inferable by the verb. In particular the 1st singular person of the verb "prendo" (*order*) indicated the dropped subject "io" (*I*) and the 3rd singular person of the verb "è stata incoronata" (*was crowned*) combined with the singular feminine suffix *-a* implies a subject pronoun of the same gender and number "Ella" (*She*).

This omissibility of the subject has been here chosen to further deepen the extent of mBERT in transferring syntactic relations against a specific phenomenon. Conversely, English and French have a syntax with much less variability of constituents and they require a mandatory explicit subject. English has quite limited inflectional variations and a rigid word order, even if compared to other Germanic languages such as German (Liu, 2010). There is a well known correlation between poor morphological variations and the degree of freedom of word order. In English the word order is so strict to avoid ambiguities caused by the lack of inflected forms. Each constituent must be expressed to indicate its syntactic function (Solodow, 2010; Vennemann, 1974; Bauer, 2009). For instance, in specific construction the so-called dummy pronoun is inserted only with syntactic function, without having any meaning (e.g. constructions like "it rains, it seems, it is important to know"). French is not a strictly morphologically rich language (Seddah et al., 2013) but it has an inflectional system richer than English and a very limited amount of word order variation occurring at different syntactic levels including the word level. These features bring it closer to Italian.

Besides a historical point of view, the proximity between the two languages has been quantitatively assessed by recent studies (Liu and Xu, 2012). Using dependency treebanks and syntactic networks the similarities between Romance languages, showing that Italian and French have a similarity degree exceeding 80%. However, French differs from most Romance languages like Italian or Spanish in a crucial aspect: during its evolution, it has reduced its word order to the single type SVO (Alexiadou, 2006). Reasons that have led to an increased syntactic rigidity with the progressive loss of pro-drop and the fixation of the subject position have been extensively discussed in the literature (Marchello-Nizia, 2006; Buridant and Zink, 2000; Lahousse and Lamiroy, 2012). This particular feature make French very interesting in comparison to English because despite its similarity to Italian (Abeillé et al., 2020; Godard, 1988) it is very close to English because of the lack of the pro-drop option.

These phenomena have been widely studied in Theoretical Linguistics (Rizzi, 1986; Gilligan, 1989; Camacho, 2013), and in the context of language comparison and language learning (Rothman, 2009), offering some interesting insights for a qualitative comparison. As highlighted by NLP studies, this syntactic misalignment produces difficulties in correctly identifying syntactic dependencies. Translation between pro-drop and non pro-drop languages has always been challenging, since translation of such missing pronouns cannot be normally reproduced (Wang et al., 2017, 2018).

[2] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip.

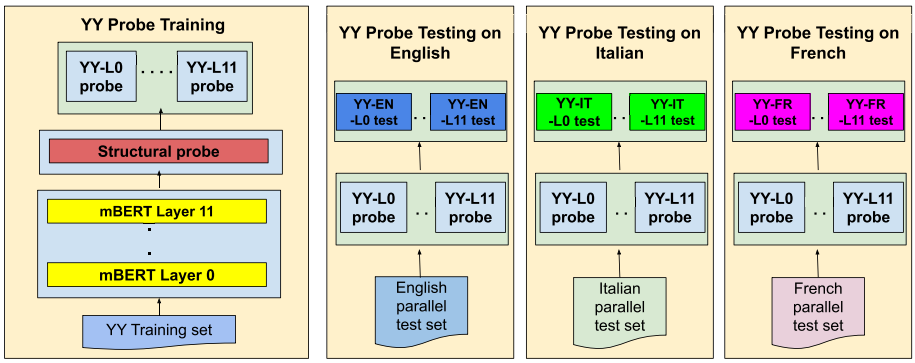


Fig. 2. Training of the probes on YY language (YY can be equal to EN, IT or FR to indicate English, Italian and French, respectively) and testing of the probes on the parallel test sets.

Furthermore, it is also important to note that the task of resolving pro-drop or null-subject phenomena has received a lot of interest in recent years. Several monolingual studies have been proposed addressing the phenomenon in even quite different languages (Chen and Ng, 2016; Ferrández and Peral, 2000; Gopal and Jha, 2017; Spence Green and Manning, 2009; Grigorova, 2013), while studies on Italian are mostly theoretical and outdated (Di Eugenio, 1996).

4. Experiments

This Section first describes the experimental assessment performed to test the cross-lingual mBERT model, then the datasets in different languages used to train and to test the probe, and finally, the metrics chosen for the evaluation.

4.1. Experimental assessment

Several experiments have been performed, with the purpose of investigating on the cross-lingual capability of the mBERT model to approximate syntactic dependency relationships in sentences and to transfer them to other languages. They have been assessed also in accordance with the results of recent studies (Tenney et al., 2019a) that have already demonstrated that monolingual BERT encodes syntax more than semantics.

Moreover, these experiments have been also aimed at determining the best layer of the Transformer stack of mBERT in embedding this syntactic information of the sentence in such a cross-lingual task. Already in Jawahar et al. (2019) different layers of a BERT-based model have shown to capture different levels of language complexity: lower layers encode phrase-level information, middle layers deal with syntactic features and higher ones focus on semantic features. Moreover, deeper layers are required in order to perform tasks related to long-distance dependency information, such as subject–verb agreement.

More in detail, the experiments have been setup as shown in Fig. 2 and described in the following. The syntactic probes leveraging each of the 12 layers of the mBERT model have been trained for all three considered languages, namely English, French and Italian, producing 36 different models denoted hereafter as *YY-LX probe*, *YY-LX probe* and *YY-LX probe*, where *YY* refers to the specific source language used (YY can be equal to EN, IT or FR to indicate English, Italian and French, respectively) and *LX* is intended to represent the specific layer chosen (X can vary from 0 to 11).

Each of these models has been then assessed on the test sets, respectively in English, French and Italian, generating 108 different experiments, denoted in the following as *YY-EN-LX test*, *YY-IT LX test* and *YY-FR-LX test*, where *YY* refers to the specific source language used for training the model and *LX* again represents the layer chosen. All experiments have been run ten times, calculating the average and the standard deviation of each metric considered.

Then, another set of experiments have been arranged with the purpose of analysing the cross-lingual capability of the mBERT model on facing the issues related to the omissibility of the subject. In particular, the previous set of experiments has been repeated on a different test set, formed only by the sentences of the test set where the subject is omitted in Italian (see Section 4.2). In this latter case, only the probes trained leveraging the best performing mBERT layers observed in the first set of experiments have been tested.

4.2. Datasets

The three datasets used for training and testing the models are morpho-syntactically annotated using Universal Dependencies (UD) v2 formalism (Nivre et al., 2020b), and they are included in the UD version 2.7 treebanks.[3] They comply with the following requirements: (i) they are robust and widely used in the literature; (ii) they are large enough to be used to train the probe; (iii) they share the same formalism for the annotation (CoNLL-U).

[3] Available at http://hdl.handle.net/11234/1-3424.

Table 2
Datasets features and sizes.

Dataset	Language	Sentence count	Average sentence length	Total word count
IT-ISDT training set	Italian	13,058	21	269,685
EN-EWT training set	English	12,543	16	204,585
FR-GSD+SEQ training sets	French	11,212	24	274,381
PUD test set	Italian	1000	24	23,731
PUD test set	English	1000	21	21,176
PUD test set	French	1000	25	24,734
PUD test set subset	Italian	120	19	2306
PUD test set subset	English	120	18	2232
PUD test set subset	French	120	21	2534

In detail, for the English language, the dataset used is *the Universal Dependencies - English Web Treebank (EN-EWT)*[4] (Silveira et al., 2014). In case of the Italian language, the training set is chosen from *Italian ISDT Treebank (IT-ISDT)*[5] (Bosco et al., 2013; Simi et al., 2014), a CoNLL–compliant Italian Treebank. For the French language two reference datasets for the French language have been merged: *UD-French-GSD (FR-GSD)*[6] (Guillaume et al., 2019) and *UD-French-Sequoia (FR-SEQ)*[7] (Candito et al., 2014). In this way, also the French training set has a comparable size to those of the other two languages under examination, as shown in the next Table 2.

Parallel sentences available in the test set of *Parallel Universal Dependencies* (PUD) (Zeman et al., 2017) have been chosen for testing the trained models. PUD is a set of parallel treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies, containing, among the others, parallel English–French–Italian sentences, manually translated and annotated. The use of this dataset, which contains the same set of annotated sentences in different languages, is mandatory to investigate the cross language capability of the mBERT model. As explained above, the structural probes trained in English, Italian and French have been respectively tested on both English, Italian and French subsets of the PUD dataset. The same sentences in different languages of these subsets have allowed the comparison of the results obtained in the cross language experiments. Moreover, they have enabled the analysis of the behaviour of the probes when they are used to obtain the DPTs of those sentences that show the considered syntactic phenomena in one language, comparing the same DPTs obtained applying the same probe on the sentences in the other languages.

All the above described datasets collect sentences coming from different domains and genres, ranging from Wikipedia articles to talks and legal texts. The sentences present a huge variability in terms of length, lexical and syntactical complexity, varying in range from 2 to 310 words. On the other hand, not all these sentences can provide interesting information, in particular on specific syntactic phenomena of the considered languages.

Focusing the analysis on the specific syntactic phenomenon of the omissibility of the subject, a further subset of sentences from the parallel PUD test set has been created. Criteria for the sentence selection are described below. First, sentences with no verb have been excluded. Then, thresholds on the minimum and maximum number of words in a sentence have been established. Recent studies (Lakretz et al., 2020) have quantitatively estimated the Syntactic Capacity Limitation by human working memory and by computational language models for the correct understanding of the syntactic complex relations of a well-formed sentence. Finally, only sentences presenting an explicit subject pronoun in Italian have been taken into account. In this way, an additional parallel Italian–English–French test set counting 120 specific sentences – whose length ranges from 3 to 40 words – has been obtained (see next Table 2).

The Table 2 summarises the languages, the total sentence count, the average sentence length and the total word count of each of the above described datasets.

4.3. Metrics

Several metrics have been defined in literature for the evaluation of dependency parsers (Buchholz and Marsi, 2006; Eisner, 1996; Kübler et al., 2009; Nivre and Fang, 2017). This experimental assessment uses two metrics related to the ***Unlabelled Attachment Score*** (UAS) metric, which is the percentage of predicted words that have the correct head. This single accuracy metric can be applied to dependency parsing thanks to the single-head property of dependency trees.

The UAS can be macro-averaged or micro-averaged on each sentence of the dataset, respectively obtaining *Word-based UAS* ($wUAS$) and *Sentence-based UAS* ($sUAS$) metrics, defined as:

- **Sentence-based UAS** $sUAS$, is the macro-averaged UAS, calculated as:

$$sUAS = \frac{1}{m}\sum_{i=1}^{m}\frac{PEdges_i}{Edges_i} \tag{3}$$

[4] https://github.com/UniversalDependencies/UD_English-EWT.
[5] https://github.com/UniversalDependencies/UD_Italian-ISDT.
[6] https://github.com/UniversalDependencies/UD_French-GSD.
[7] https://github.com/UniversalDependencies/UD_French-Sequoia.

where m is the total number of sentences in the dataset, $PEdges_i$ is the number of correctly predicted edges for the ith sentence and $Edges_i$ is the total number of true edges of the ith sentence.

- **Word-based UAS** $wUAS$, is calculated as the fraction of the correctly predicted edges over the total number of edges among the words of all sentences of the dataset, as:

$$wUAS = \frac{PEdges}{Edges} \tag{4}$$

where $PEdges$ is the total number of correctly predicted edges and $Edges$ is the total number of true edges of the whole dataset.

The $sUAS$ is an average of the UAS calculated on each single sentence of the dataset, and its value is less affected by outliers related to long sentences, where a higher percentage of errors can occur. On the other hand, the $wUAS$ can provide a more general index of the performances of the model, taking into account only the correct edges, despite their respective sentence.

4.4. Experimental setup

The probes have been trained using a batch size and a number of epochs respectively set to 20 and to 40, following the experiments described in Hewitt and Manning (2019), leveraging the embeddings extracted from the hidden states of the mBERT model described in previous Section 3.2.

The experiments have been run on an IBM Power9-based system, a cluster computing hardware specifically devoted to deep learning, counting by nodes each one with two Power9 CPUs clocked at 3.7 GHz, with 512 GB of RAM and with four Nvidia Tesla V100 GPUs with 16 GB of dedicated VRAM. The operating system of the cluster is Red Hat Enterprise Linux Server release 7.6. Using a single GPU, the extraction of contextual word embeddings from each layer of the mBERT architecture has required approximately 250 s every 25,000 words, while the training of the structural probe has required an average run time equal to about 200 s each 5000 sentences of the training set.

5. Results and discussion

In this Section the results of the experimental assessment aiming at investigate the cross-lingual syntactic transfer capabilities of mBERT are first presented and discussed (Section 5.1). Then, the cross-lingual transfer in case of omissibility of the subject syntactic phenomenon is analysed and discussed, from both a quantitative and a qualitative point of view (Section 5.2).

5.1. Cross-lingual syntactic transfer results

The next Tables 3–5 show the average and the standard deviation of the $sUAS$ and $wUAS$ metrics obtained for the experiments performed on the whole PUD test sets with each probe exploiting a different mBERT layer and respectively trained in English, Italian and French. In addition, Fig. 3 shows a plot of the results of the Tables 3–5, providing a more compact visualisation of the metrics and allowing for a global comparison of their trends.

Observing the global behaviour of the probes in all the Tables, it is first possible to confirm that, in case of mBERT model, the layers where more syntactic information is embedded are the central upper ones for cross-lingual experiments, as previously observed in monolingual English BERT-Base model (Jawahar et al., 2019; Hewitt and Manning, 2019). In particular, the probes based on mBERT and trained on English (Table 3) have produced the best results at layer 6 with EN-EN and EN-FR tests and at layer 7 with EN-IT test. The probes trained in Italian (Table 4) have produced the best results for both monolingual and cross-lingual tests, respectively IT-IT, IT-EN and IT-FR tests, exploiting the embeddings from the layer 6 of mBERT. Finally, the probes trained in French (Table 5) have obtained the best results for $sUAS$ respectively at the layer 7 for FR-FR and FR-EN tests and at the layer 6 for FR-IT test, while the $wUAS$ has produced best results at layer 6 for FR-FR and FR-EN tests and at layer 7 for FR-IT test. Previous experiments for monolingual English described in Hewitt and Manning (2019) have showed that the same probe exploiting English BERT-Base model has produced the best results at layer 7 in term of $wUAS$.

5.1.1. Discussion

Observing the metrics in the first two columns of the Tables (monolingual experiments, in the first two columns of Tables 3–5) and the ones in the other four columns (cross-lingual experiments), it is possible to note that the results in cross-lingual tests are comparable with the monolingual ones. It is worth noting that in this latter case the probes achieved metrics comparable with the ones previously described in literature (Jawahar et al., 2019; Hewitt and Manning, 2019).

In general, the observed behaviour demonstrates that the capability of mBERT model to embed the sentence structures is not affected by the specific languages, demonstrating its ability in cross-lingual syntax transfer.

As expected, the cross lingual tests have showed a slight performance drop, whose differences among various cases are in accordance with the Theoretical Linguistics, as explained below.

As shown in Table 3, English-trained probes have achieved good scores both on French and Italian. In detail, they have shown slight better results in EN-FR test. This good approximation of syntax in French using English is not surprising (see last columns of Table 3), because it coincides with well known motivations in Theoretical Linguistics. Indeed, for historical reasons, they both share some of the syntactical properties. For instance, they are languages without grammatical cases and they do not admit the absence

Table 3
Results for each EN-LX probe respectively tested on PUD English, Italian and French test sets.

mBERT layer	sUAS (English test set) EN-EN	wUAS (English test set) EN-EN	sUAS (Italian test set - Cross Lingual) EN-IT	wUAS (Italian test set - Cross Lingual) EN-IT	sUAS (French test set - Cross Lingual) EN-FR	wUAS (French test set - Cross Lingual) EN-FR
0	0.5191 ± 0.0043	0.5079 ± 0.0030	0.4167 ± 0.0021	0.4044 ± 0.0023	0.4211 ± 0.0059	0.4168 ± 0.0050
1	0.5791 ± 0.0026	0.4564 ± 0.0040	0.5011 ± 0.0074	0.4877 ± 0.0080	0.4971 ± 0.0068	0.4927 ± 0.0060
2	0.6782 ± 0.0055	0.6678 ± 0.0054	0.4976 ± 0.0078	0.4865 ± 0.0067	0.5719 ± 0.0056	0.5628 ± 0.0044
3	0.7137 ± 0.0048	0.6993 ± 0.0046	0.6280 ± 0.0059	0.6161 ± 0.0046	0.6275 ± 0.0047	0.6187 ± 0.0032
4	0.7435 ± 0.0038	0.7312 ± 0.0039	0.6529 ± 0.0061	0.6400 ± 0.0042	0.6446 ± 0.0042	0.6325 ± 0.0055
5	0.7740 ± 0.0035	0.7589 ± 0.0033	0.6815 ± 0.0072	0.6646 ± 0.0060	0.6827 ± 0.0043	0.6699 ± 0.0020
6	**0.7927 ± 0.0035**	**0.7909 ± 0.0032**	0.6825 ± 0.0044	0.6667 ± 0.0036	**0.7094 ± 0.0017**	**0.6899 ± 0.0014**
7	0.7837 ± 0.0047	0.7633 ± 0.0040	**0.6903 ± 0.0051**	**0.6667 ± 0.0029**	0.7022 ± 0.0049	0.6780 ± 0.0050
8	0.7614 ± 0.0032	0.7444 ± 0.0028	0.6726 ± 0.0037	0.6602 ± 0.0034	0.6862 ± 0.0044	0.6706 ± 0.0047
9	0.7414 ± 0.0053	0.7248 ± 0.0043	0.6166 ± 0.0046	0.6047 ± 0.0033	0.6576 ± 0.0076	0.6444 ± 0.0076
10	0.7320 ± 0.0035	0.7174 ± 0.0031	0.5852 ± 0.0066	0.5737 ± 0.0057	0.6335 ± 0.0043	0.6210 ± 0.0052
11	0.6629 ± 0.0047	0.6444 ± 0.0039	0.5659 ± 0.0050	0.5471 ± 0.0043	0.5524 ± 0.0068	0.5400 ± 0.0054

Table 4
Results for each IT-LX probe respectively tested on the Italian, English and French PUD test sets.

mBERT layer	sUAS (Italian test set) IT-IT	wUAS (Italian test set) IT-IT	sUAS (English test set - Cross Lingual) IT-EN	wUAS (English test set - Cross Lingual) IT-EN	sUAS (French test set - Cross Lingual) IT-FR	wUAS (French test set - Cross Lingual) IT-FR
0	0.5891 ± 0.0045	0.5782 ± 0.0043	0.3974 ± 0.0038	0.3919 ± 0.0024	0.4646 ± 0.0034	0.4593 ± 0.0043
1	0.4480 ± 0.0013	0.4479 ± 0.0012	0.4424 ± 0.0035	0.3987 ± 0.0032	0.5433 ± 0.0039	0.5413 ± 0.0043
2	0.7143 ± 0.0038	0.7054 ± 0.0039	0.5119 ± 0.0021	0.5095 ± 0.0033	0.6303 ± 0.0052	0.6238 ± 0.0055
3	0.7506 ± 0.0046	0.7434 ± 0.0047	0.5662 ± 0.0024	0.5554 ± 0.0024	0.6761 ± 0.0053	0.6733 ± 0.0038
4	0.7623 ± 0.0049	0.7555 ± 0.0050	0.6184 ± 0.0072	0.6077 ± 0.0060	0.7103 ± 0.0023	0.7024 ± 0.0017
5	0.7840 ± 0.0039	0.7776 ± 0.0048	0.6405 ± 0.0070	0.6314 ± 0.0064	0.7327 ± 0.0024	0.7211 ± 0.0027
6	**0.7972 ± 0.0020**	**0.7875 ± 0.0032**	**0.6672 ± 0.0011**	**0.6483 ± 0.0017**	**0.7418 ± 0.0051**	**0.7226 ± 0.0050**
7	0.7874 ± 0.0019	0.7711 ± 0.0024	0.6529 ± 0.0066	0.6349 ± 0.0062	0.7395 ± 0.0016	0.7192 ± 0.0012
8	0.7731 ± 0.0031	0.7662 ± 0.0032	0.6505 ± 0.0037	0.6346 ± 0.0033	0.7257 ± 0.0029	0.7124 ± 0.0038
9	0.7543 ± 0.039	0.7485 ± 0.0048	0.6346 ± 0.0026	0.6209 ± 0.0028	0.6903 ± 0.0029	0.6761 ± 0.0027
10	0.7487 ± 0.0015	0.7453 ± 0.0009	0.6053 ± 0.0024	0.5931 ± 0.0015	0.6788 ± 0.0055	0.6635 ± 0.0049
11	0.7069 ± 0.0051	0.6979 ± 0.0063	0.5147 ± 0.0083	0.4065 ± 0.0069	0.5900 ± 0.0062	0.5773 ± 0.0043

Table 5
Results for each FR-LX probe respectively tested on the French, English and Italian PUD test sets.

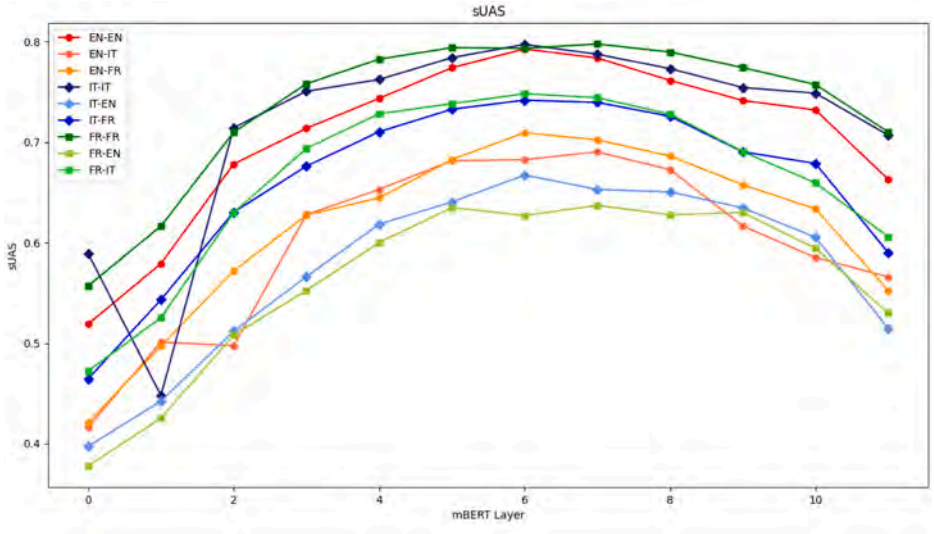
mBERT layer	sUAS (French test set) FR-FR	wUAS (French test set) FR-FR	sUAS (English test set - Cross Lingual) FR-EN	wUAS (English test set - Cross Lingual) FR-EN	sUAS (Italian test set - Cross Lingual) FR-IT	wUAS (Italian test set - Cross Lingual) FR-IT
0	0.5571 ± 0.0029	0.5508 ± 0.0022	0.3778 ± 0.0053	0.3697 ± 0.0034	0.4724 ± 0.0044	0.4587 ± 0.0043
1	0.6165 ± 0.0022	0.6140 ± 0.0026	0.4255 ± 0.0057	0.4175 ± 0.0077	0.5253 ± 0.0045	0.5142 ± 0.0046
2	0.7098 ± 0.0061	0.6969 ± 0.0053	0.5082 ± 0.0065	0.4955 ± 0.0049	0.6298 ± 0.0031	0.6179 ± 0.0035
3	0.7580 ± 0.0040	0.7437 ± 0.0039	0.5523 ± 0.0067	0.5438 ± 0.0058	0.6937 ± 0.0051	0.6832 ± 0.0053
4	0.7825 ± 0.0054	0.7694 ± 0.0039	0.6001 ± 0.0067	0.5804 ± 0.0062	0.7282 ± 0.0040	0.7174 ± 0.0027
5	0.7940 ± 0.0024	0.7812 ± 0.0027	0.6348 ± 0.0036	0.6191 ± 0.0043	0.7383 ± 0.0037	0.7287 ± 0.0037
6	0.7933 ± 0.0030	**0.7817± 0.0020**	0.6271 ± 0.0037	0.6180 ± 0.0027	**0.7481 ± 0.0056**	0.7336 ± 0.0053
7	**0.7976 ± 0.0024**	0.7779 ± 0.0025	**0.6370 ± 0.0050**	**0.6208 ± 0.0052**	0.7442 ± 0.0019	**0.7346 ± 0.0026**
8	0.7896 ± 0.0020	0.7719 ± 0.0020	0.6278 ± 0.0033	0.6083 ± 0.0040	0.7278 ± 0.0040	0.7237 ± 0.0047
9	0.7741 ± 0.0052	0.7563 ± 0.0046	0.6300 ± 0.0060	0.6132 ± 0.0042	0.6906 ± 0.0051	0.6840 ± 0.0057
10	0.7573 ± 0.0031	0.7435 ± 0.0026	0.5944 ± 0.0028	0.5789 ± 0.0020	0.6594 ± 0.0015	0.6507 ± 0.0018
11	0.7100 ± 0.0016	0.6919 ± 0.0026	0.5299 ± 0.0101	0.5097 ± 0.0098	0.6059 ± 0.0112	0.5937 ± 0.0104

of the subject pronoun. In addition, they show a limited degree of freedom in the order of the constituents (i.e. they do not allow a good part of the possible movements of syntactic arguments in Italian).

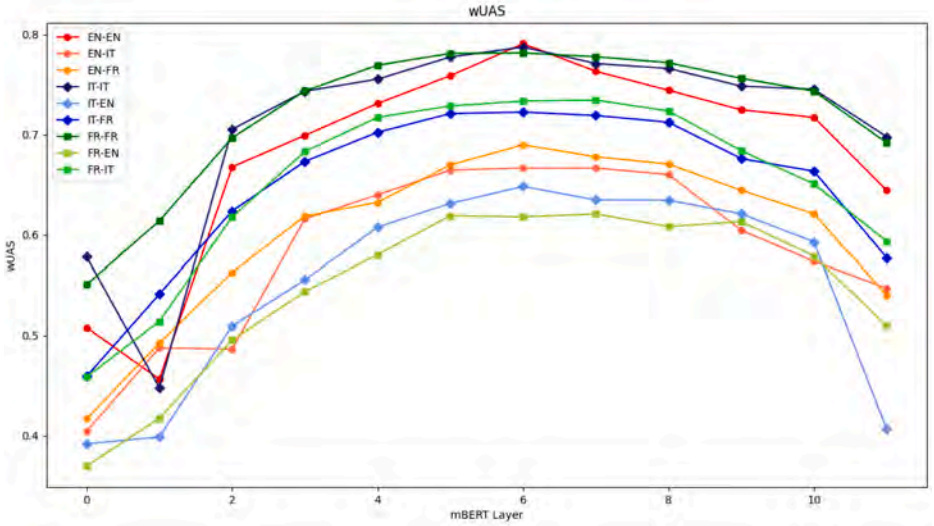
As expected, transferring English syntax on Italian has resulted slightly more complicated, with a difference of about one percentage point compared to French case (see central columns of Table 3). This could be partly attributed to a fundamental difference between English and Italian, that belong to different branches of the Indo-European language family (respectively Germanic and Italic languages) with several differences on syntactic constructions. As further confirmation, this behaviour has been also maintained on the opposite example, namely the IT-EN experiment (as shown in Table 4).

The behaviour of Italian trained probe also has a historical–theoretical interpretation. As expected, the results of IT-FR experiments in Table 4 have shown that the performances are higher than IT-EN ones. This can be easily explained by the syntactic and lexical similarities of Italian and French, which have a common matrix: they both belong to Romance languages, a sub-group of Italic languages in the Indo-European family. As an additional evidence, training the model on French has given excellent results on the Italian language (as shown in Table 5).

Finally, French trained model has worked differently if applied to Italian or English (Table 5). FR-IT test has achieved a good score, while FR-EN test has gotten lower results. It is important to discuss the specific case of the syntax transfer from French to



(a) sUAS



(b) wUAS

Fig. 3. Plots of the results of obtained in all experiments in terms of sUAS and wUAS.

English. As already observed above, the model trained on English has been able to transfer the syntax to French with excellent results (see last two columns of Table 3), but this behaviour has not been observed in the opposite case.

This suggests that the common characteristics between English and French only partially affect syntax learning. Although French shares many syntactical rules and constraints with English (i.e. fixed structure, mandatory express subject), it has a more sophisticated articulation of some constructions (e.g. passive structures, long-distance dependencies, negative form, relative clause, etc.). These features represent a legacy of its historical closeness to the Italian syntax and to the group of other Romance languages. This allows the model trained on French to learn with good approximation the syntax of Italian, while it faces significant difficulties trying to approximate syntactic structure non-existent in the target language, as in the case of FR-EN tests.

5.2. Subject omission results

The Tables 6–8 present the results obtained by the probes in the cross-lingual experiments, focusing only on the selected subset of the test set (see previous Section 4.2) where the corresponding Italian sentences show the linguistic phenomenon of the *omitted*

Table 6
Results for the best performing EN probes (EN-6 and EN-7) respectively tested in English, Italian and French on a subset of sentences where the subject is omitted in Italian.

mBERT layer	sUAS (English test set) EN-EN	wUAS (English test set) EN-EN	sUAS (Italian test set - Cross Lingual) EN-IT	wUAS (Italian test set - Cross Lingual) EN-IT	sUAS (French test set - Cross Lingual) EN-FR	wUAS (French test set - Cross Lingual) EN-FR
6	**0.8155 ± 0.0030**	**0.7893 ± 0.0040**	**0.7151 ± 0.0042**	**0.6960 ± 0.0052**	**0.7148 ± 0.0054**	**0.6905 ± 0.0060**
7	0.7994 ± 0.0028	0.7694 ± 0.0036	0.7149 ± 0.0064	0.6935 ± 0.0053	0.7066 ± 0.0035	0.6819 ± 0.0026

Table 7
Results for the best performing IT probes (IT-6 and IT-7) respectively tested in Italian, English and French on a subset of sentences where the subject is omitted in Italian.

mBERT layer	sUAS (Italian test set) IT-IT	wUAS (Italian test set) IT-IT	sUAS (English test set - Cross Lingual) IT-EN	wUAS (English test set - Cross Lingual) IT-EN	sUAS (French test set - Cross Lingual) IT-FR	wUAS (French test set - Cross Lingual) IT-FR
6	**0.8290 ± 0.0024**	**0.8112 ± 0.0027**	**0.6813 ± 0.0071**	**0.6574 ± 0.0075**	**0.7524 ± 0.0020**	**0.7325 ± 0.024**
7	0.8158 ± 0.0019	0.7898 ± 0.0022	0.6730 ± 0.0043	0.6485 ± 0.0029	0.7376 ± 0.0052	0.7147 ± 0.0056

Table 8
Results for the best performing FR probes (FR-6 and FR-7) respectively tested in French, English and Italian on a subset of sentences where the subject is omitted in Italian.

mBERT layer	sUAS (French test set) FR-FR	wUAS (French test set) FR-FR	sUAS (English test set - Cross Lingual) FR-ENG	wUAS (English test set - Cross Lingual) FR-ENG	sUAS (Italian test set - Cross Lingual) FR-IT	wUAS (Italian test set - Cross Lingual) FR-IT
6	**0.8209 ± 0.0017**	**0.8021 ± 0.0027**	**0.6655 ± 0.0020**	**0.6473 ± 0.0030**	**0.7808 ± 0.0029**	**0.7650 ± 0.0020**
7	0.8067 ± 0.0014	0.7876 ± 0.0023	0.6616 ± 0.0024	0.6396 ± 0.0029	0.7743 ± 0.0049	0.7556 ± 0.0042

Table 9
Percentage of arcs among subject and corresponding verb correctly predicted in the English and French subsets of sentences where the subject is omitted in Italian in case of probe using embeddings from layer 6 of mBERT trained in Italian (columns 2 and 3), compared with the percentage of the same correctly predicted arcs in the case of monolingual experiments (EN-EN and FR-FR).

Layer	IT-EN	IT-FR	EN-EN	FR-FR
6	69.17%	82.50%	75.83%	84.17%

subject (pro-drop). Following the previous results, this further analysis has been focused only on the two best performing layers (6 and 7) of the mBERT model.

Analysing the results of these further experiments, it is possible to first note a global performance improvement in terms of both $sUAS$ and $wUAS$. This performance boost in experiments can be partially attributed to the higher number of well-formed and less complex sentences. The results in Tables 6–8 have proved that the mBERT model is able to encode this particular phenomenon when exploited in a cross-lingual task, being these metrics comparable with the ones obtained in the previous general case shown in Tables 3–5.

To the end of analysing more in detail the cross language transfer of the subject omission phenomenon from Italian to the other languages, the next Table 9 shows the percentage of the correctly predicted arcs between subject and verb in English and French sentences whose corresponding parallel Italian sentences show the pro-drop. In particular, the second and the third columns of Table 9 show this percentage obtained by the probe exploiting the embeddings from the layer 6 of the mBERT model and trained in Italian, respectively tested on English and French subset of the test set. The last two columns show the percentage of correctly predicted arcs between subject and verb for the monolingual English and French experiments (EN-EN and FR-FR).

Also in this case, the results obtained with the probes trained in Italian and tested in English and French are comparable with the ones obtained in the corresponding monolingual experiments, providing further evidences that the mBERT model is able to embed the knowledge of the subject agreement, although the languages used to train the probe allows for pro-drop phenomenon. It is worth noting that in the case of French test set, the probes respectively trained in Italian and in French have produced almost similar results, probably also thanks to the high level of similarity among these two languages.

The following qualitative analysis can better show the behaviour of the probe in this case.

5.2.1. Qualitative analysis and discussion

As examples, two sentences in Figs. 4 and 5 have been chosen since satisfying two main requirements. First, they do not present great lexical variability from one language to another. Please note that the PUD corpus does not contain literal word-for-word translations, but adapted to the syntax and vocabulary of each language. Secondly, they allow to observe different syntactic complexity: the first one contains an adversative structure, while the second one is more complex, with a relative clause.

The Figures are structured as follows. In the left column, the three versions of the sentence in the different languages are reported with the syntactic dependencies obtained from the monolingual test. Dependencies are shown by the corresponding arcs (red for

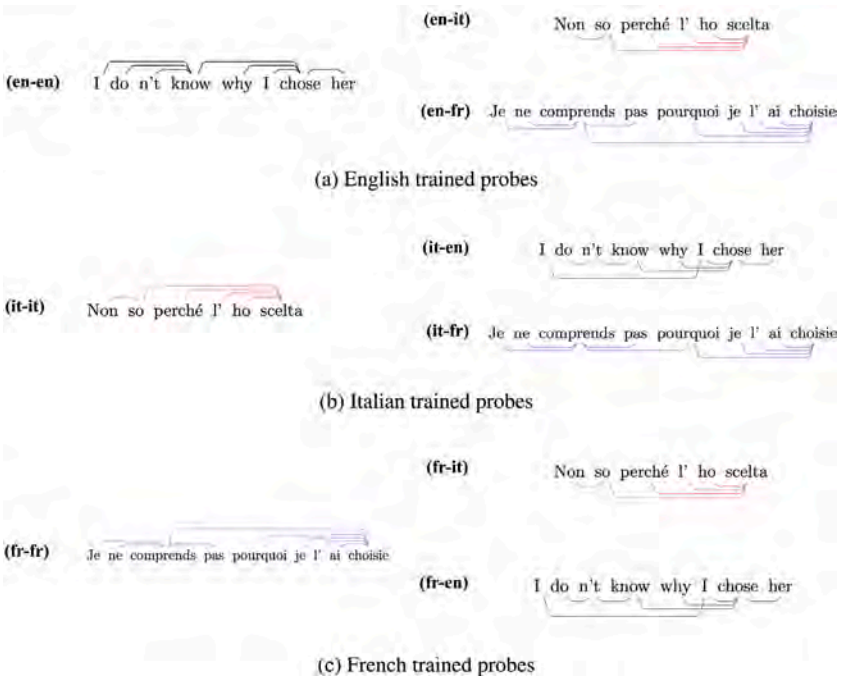


Fig. 4. Example of cross-lingual realisations for a sentence with a adversative construction expressed by a negative clause. On the left, dependencies created using EN-EN-L6, IT-IT-L6 and FR-FR-L6 probes, on the right, the ones created cross-lingually on the target language. Colours distinguish different languages (red for Italian, black for English and blue for French). Monolingual arcs are shown above the sentences, arcs cross-lingual approximated are shown below. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

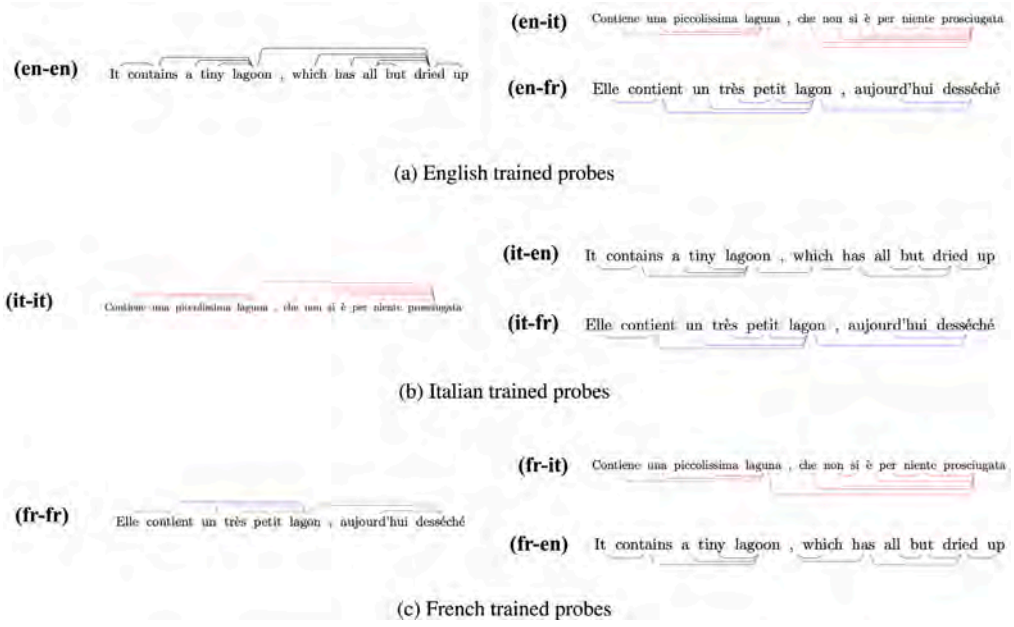


Fig. 5. Example of cross-lingual realisations for a dual-verb sentence with a relative clause and a long-distance dependency. On the left, dependencies created using EN-EN-L6, IT-IT-L6 and FR-FR-L6 probes, on the right, the ones created cross-lingually on the target language. Colours distinguish different languages (red for Italian, black for English and blue for French). Monolingual arcs are shown above the sentences, arcs cross-lingual approximated are shown below. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Italian, black for English and blue for French). The right column shows dependencies created on every target language with the cross-lingual experiments, exploiting the respective best performing layer of mBERT.

The sentence in Fig. 4 presents a principal clause with adversative value followed by a causal proposition. It is important to note that the structure of the negative sentence has different constructions in each language. Italian puts a negative adverb *"non"* before the verb, which gives negative value to the concept expressed by the verb to which it is premised *"so"*. This latter is followed by *"perché"*, which expresses the motivation for the negation. English requires the construction with the auxiliary *"do not"*, while French construct for negation is made of the particle *"ne"* before the verb and the *"pas"* immediately after. In the Italian sentence there is no explicit subject. All information is taken from the inflected form of the verb. Corresponding sentences in English and French respectively express the subject with singular first-person pronouns *"I"* and *"Je"*.

Concerning the model trained on English as a source language (see Fig. 4a), a similar behaviour has occurred in the other languages, in line with the quantitative results.

Syntax transfer has been slightly worse on Italian than on French, even if with a minimal gap. Although all EN-EN dependencies are correct, it is possible to assume that the drop in performance is a result of the different structure of the causative clause. Indeed, while in English the object pronoun immediately follows the verb of the subordinate clause *I chose her*, both in Italian and French the pronoun precedes it: *"l'ho scelta"* and *"je l'ai choise"*.

IT-EN experiments have shown poor performances (right column of Fig. 4b), probably due to the fact that the subject of the main *"I"* has been not related to the negative structure of the main sentence. The principal verb *"know"* has a first relation with the subject of the clause *"I"*, and a second relation the verb *"chose"*. Remaining relations have been correctly maintained. On the other hand, IT-FR experiments have achieved a good precision, consistent with the scores reported. The verb of the main clause *"comprends"* has been correctly connected to the verb of the subordinate clause *"ai choise"* as well as the other relations have been correctly learnt. Probably the difference in performance between IT-EN and IT-FR is simply due to the change of the negative structure, which is different for both French and English.

Finally, tests carried out using a model trained for the French language have been taken into account (see Fig. 4c). The reconstructed relations for the Italian language have been almost all correct. In particular, the link between verb and subject (unexpressed in Italian) has been correctly approximated for the target language, eliminating the arc that in the French sentence connects *"Je"* to *"comprends"*, without affecting the other dependencies. This precision has been not maintained on the English language, although the syntactic structure with the explicit subject *"I"* is similar. Most of the dependencies have been erroneous. In particular, the subject of the main clause *"I"* has been related to the subject of the subordinate clause *"I"* and not to its verb *"know"*; the main verb has been related only to the negative conjunction *"not"*.

The second sentence in Fig. 5 has a more complex syntactic structure, involving a main construction followed by relative clause introduced by a pronoun. As shown in Table 6, the EN-EN experiments (Fig. 5a) have shown equivalence of performance if applied to Italian and French. The syntactic relations have been correctly preserved, with the single exception of the subject expressed in French using the third person singular *"Elle"*.

Using Italian as the source language the better approximated syntactic dependencies have been the ones for the French language (Fig. 5b). Actually, the syntactic tree reconstructed in French has maintained all the relations unchanged. On the contrary, English has achieved much lower performances. In English, the relations between the object and the verb have been inverted. The noun with object function *"lagoon"* has been not connected to the verb *"has dried up"* but to the relative pronoun *"which"* that introduces the subordinate clause.

FR-IT experiments (Fig. 5c) have achieved significantly better performance than FR-EN. FR-IT tests have correctly maintained all relations both within the main and subordinate clauses. Conversely, reconstructed dependencies for English have been strongly altered. Even dependencies of the main verb have been wrong, for instance the object of the main verb *"lagoon"* has no arc related to its verb *"has dried up"*.

Summarising, the tests have shown a better learning capacity between Italian and French, English with Italian and French as target languages and from French to Italian. It is worth noting that, given the preliminary nature of this study, there can be no claim of representativeness, both with regard to the dataset examined and the single syntactic phenomenon analysed.

Some observations can be made in relation to the complexity of the sentences. In the case of a simple sentence as for the first example (Fig. 4), most problematic relations have been those between the verb and its argument, in particular they have worsened in conjunction with an increase of the distance between these elements. This has happened because of the occurrence of syntactic elements typical of English and French construction, instead not present in Italian.

Notice that, although there is no single metric for classifying the complexity of one language compared to another (McWhorter, 2001; Ferguson, 1982), at syntactic level word order freedom and pro-drop parameter are indicators that significantly increase the complexity of a language (Brunato and Dell'Orletta, 2017). A further confirmation is given by the greater difficulty of parsing in languages with greater freedom of constituents (Seddah et al., 2013; Tsarfaty et al., 2012).

This phenomenon is noticeable in the case of the second sentence (see Fig. 5) which has two verbs and a relative sentence introduced by *"che"* in Italian. In general, it is possible to hypothesise that learning becomes more difficult in the case of unbounded distance constructions, which can include relative clauses. All the listed phenomena involve the dislocation of a constituent (filler), which is no longer *in situ* but in another position (gap). Phenomena of this type are therefore a problem in the syntactic analysis of fixed-order languages (English and French), where changes in the order of the constituents must be taken into account, and in particular they are problematic for those formalisms strongly based on linear order.

Although only three languages have been considered into the analysis, the ability of the probe to reconstruct cross-lingual dependencies – even in the presence of specific syntactic phenomena that may differ from one language to another – could bring practical benefits and implications in many NLP tasks. For instance, the knowledge of the layers where the model embeds more linguistic information and the capability in dealing with the null-subject phenomenon can be exploited to improve previous

approaches for the definition of a dependency parser, addressing some well-known open issues related to the correct parsing of null-subject phenomenon sentences in specific languages (Bosco et al., 2013; Chung et al., 2010), as well as in case of other particular syntactic phenomena, such as zero-pronoun construction (Liu et al., 2017; Song et al., 2020) or flexible clause structures (De Santo, 2019).

A clarification must be made regarding the criterion used for the choice of sentences. As already mentioned above, the work presented here is essentially corpus-based, so all the sentences are extracted from existing resources. There is still an open discussion in Linguistics about the source of the data to be used, especially in correlation with the latest models based on deep learning (Linzen, 2019).

An approach based on using naturally-occurring sentences extracted from corpora is affected by sentences that tend most probably to have simple syntactic structures (Linzen et al., 2016; Kuncoro et al., 2018). Moreover it is very difficult already on monolingual studies to find a high number of sentences showing very specific phenomena. Alternative approaches have proposed to adapt human psycholinguistic paradigms to neural networks (Marvin and Linzen, 2019). A controlled dataset of sentences specially constructed to show all syntactic phenomena to be analysed at all levels of complexity has been created to test linguistic knowledge of the network.

However, this approach is extremely expensive and unfeasible in low-resource languages. Moreover, Moreover, using corpora as resources is the only possible and consistent approach with the purpose of this work, since the crucial point of the methodology is to exploit existing resources to cross-linguistically train a model and improve performance in languages which do not have the same amount of resources as English.

6. Conclusion

This paper has presented an analysis on the capability of mBERT to transfer syntactic information about sentence structure embedded into its layers to another language, without the model being specifically trained on it. To this end, a structural probe has been exploited, demonstrating that mBERT is able to embed the dependency parse trees of the sentences cross-lingually, in particular by considering English, Italian and French.

Furthermore, the analysis has been focused on the specific null-subject phenomenon, which is peculiar of Italian language and not present in English and French. In this case, mBERT model has demonstrated its ability to reconstruct the dependency parse trees of the Italian sentences without the subject, when trained in English or French, with performances comparable to the monolingual cases. In addition, mBERT model has resulted equally able to generate the correct correspondence of the English and French sentences' subject in the trees when trained in Italian, as well as in the case of French training and English or Italian testing.

Future works plan to test the different multilingual Transformer-based NLMs, such as XLM (Conneau et al., 2020a) and different languages, with their specific syntactic phenomena. It could be also very interesting to test the behaviour of the probe when the contextual embeddings are obtained through a combination of more layers from the mBERT model, taking also into account that, observing the obtained results, the best performing NLM layers slightly varied across the different experiments performed.

From a more strictly linguistic point of view, a future development is intended to extend the analysis not only to the null-subject phenomenon, but including other peculiar aspects of the languages, such as gender/number agreement, relative clauses, passive transformations. From a more strictly linguistic point of view, a future development is intended to extend the analysis to other aspects of languages, such as gender/number agreement, relative clauses, passive transformations. As proposed by recent studies (Linzen, 2019), these analyses could benefit from the integration of materials from psycholinguistic paradigms, in order to better evaluate the extent to which the language model learns syntactic information. Other *ad-hoc* created sentences based on controlled experimental approach can be added to the existing naturally occurring corpus-based ones to increase the robustness of the methodology and highlight critical examples, adapting the methodology originally created to evaluate human ability to perform specific syntactic tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been partially supported by the Italian project "IDEHA - Innovation for Data Elaboration in Heritage Areas" funded by PON "Ricerca e Innovazione" 2014–2020.

References

Abeillé, A., Hemforth, B., Winckel, E., Gibson, E., 2020. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. Cognition 204, 104293.

Alexiadou, A., 2006. On the properties of VSO and VOS orders in greek and Italian: A study on the syntax information structure interface. In: ITRW on Experimental Linguistics. Athens, Greece, pp. 1–8. http://dx.doi.org/10.36505/ExLing-2006/01/0001/000001.

Alicante, A., Bosco, C., Corazza, A., Lavelli, A., 2012. A treebank-based study on the influence of Italian word order on parsing performance. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp. 1985–1992, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/561_Paper.pdf.

Bates, E., 1974. Language and Context: Studies in the Acquisition of Pragmatics (Ph.D. thesis). University of Chicago, Committee on Human Development.
Bates, E., McNew, S., MacWhinney, B., Devescovi, A., Smith, S., 1982. Functional constraints on sentence processing: A cross-linguistic study. Cognition 11 (3), 245–299. http://dx.doi.org/10.1016/0010-0277(82)90017-8, URL https://www.sciencedirect.com/science/article/pii/0010027782900178.
Bauer, B., 2009. Word order. New Perspect. Hist. Lat. Syntax 1, 241–316.
Bjerva, J., Augenstein, I., 2021. Does typological blinding impede cross-lingual sharing?. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp. 480–486, URL https://www.aclweb.org/anthology/2021.eacl-main.38.
Blake, B.B., 1988. Basic word order. Functional principles. J. Linguist. 24 (1), 213–217. http://dx.doi.org/10.1017/S0022226700011646.
Bosco, C., Montemagni, S., Simi, M., 2013. Converting Italian treebanks: Towards an Italian stanford dependency treebank. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. ACL, Sofia, Bulgaria, pp. 61–69, URL https://www.aclweb.org/anthology/W13-2308.
Brunato, D., Dell'Orletta, F., 2017. On the order of words in Italian: a study on genre vs complexity. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pp. 25–31.
Buchholz, S., Marsi, E., 2006. CoNLL-x shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). ACL, New York City, pp. 149–164, URL https://www.aclweb.org/anthology/W06-2920.
Buridant, C., Zink, M., 2000. Grammaire Nouvelle de L'Ancien Français. SEDES.
Burzio, L., 1986. Italian Syntax: A Government-Binding Approach, Vol. 1. Springer Science & Business Media.
Camacho, J., 2013. Null Subjects. Cambridge University Press, http://dx.doi.org/10.1017/CBO9781139524407.
Camacho-Collados, J., Pilehvar, M.T., Navigli, R., 2015. A unified multilingual semantic representation of concepts. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pp. 741–751. http://dx.doi.org/10.3115/v1/P15-1072, URL https://www.aclweb.org/anthology/P15-1072.
Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., de la Clergerie, É.V., 2014. Deep syntax annotation of the sequoia french treebank. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 2298–2305, URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/494.html.
Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., Esposito, M., 2020. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Appl. Soft Comput. 97, 106779. http://dx.doi.org/10.1016/j.asoc.2020.106779, URL http://www.sciencedirect.com/science/article/pii/S1568494620307171.
Chen, C., Ng, V., 2016. Chinese zero pronoun resolution with deep neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 778–788.
Chi, E.A., Hewitt, J., Manning, C.D., 2020. Finding universal grammatical relations in multilingual BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 5564–5577. http://dx.doi.org/10.18653/v1/2020.acl-main.493, URL https://www.aclweb.org/anthology/2020.acl-main.493.
Chomsky, N., 1957. Syntactic Structures. Mouton de Gruyter, Berlin, Germany.
Chomsky, N., 1981. Lectures on government and binding (dordrecht: Foris). Stud. Gener. Gramm. 9.
Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Chung, T., Post, M., Gildea, D., 2010. Factors affecting the accuracy of Korean parsing. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages. Association for Computational Linguistics, Los Angeles, CA, USA, pp. 49–57, URL https://www.aclweb.org/anthology/W10-1406.
Clark, K., Khandelwal, U., Levy, O., Manning, C.D., 2019. What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. ACL, Florence, Italy, pp. 276–286. http://dx.doi.org/10.18653/v1/W19-4828, URL https://www.aclweb.org/anthology/W19-4828.
Comrie, B., 1989. Language Universals and Linguistic Typology: Syntax and Morphology. University of Chicago press.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020a. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, Online, pp. 8440–8451, URL https://www.aclweb.org/anthology/2020.acl-main.747.
Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M., 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 2126–2136. http://dx.doi.org/10.18653/v1/P18-1198, URL https://www.aclweb.org/anthology/P18-1198.
Conneau, A., Lample, G., 2019. Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019. Vancouver, BC, Canada, pp. 7057–7067, URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.
Conneau, A., Wu, S., Li, H., Zettlemoyer, L., Stoyanov, V., 2020b. Emerging cross-lingual structure in pretrained language models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 6022–6034. http://dx.doi.org/10.18653/v1/2020.acl-main.536, URL https://www.aclweb.org/anthology/2020.acl-main.536.
Croft, W., 2009. Methods for finding language universals in syntax. In: Scalise, S., Magni, E., Bisetto, A. (Eds.), Universals of Language Today. Springer Netherlands, Dordrecht, pp. 145–164. http://dx.doi.org/10.1007/978-1-4020-8825-4_8.
Cruz, A., Rocha, G., Cardoso, H.L., 2018. Exploring spanish corpora for portuguese coreference resolution. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, pp. 290–295. http://dx.doi.org/10.1109/SNAMS.2018.8554705.
Davis, F., van Schijndel, M., 2020. Recurrent neural network language models always learn english-like relative clause attachment. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, pp. 1979–1990, URL https://www.aclweb.org/anthology/2020.acl-main.179.
De Santo, A., 2019. Testing a minimalist grammar parser on Italian relative clause asymmetries. In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. ACL, Minneapolis, Minnesota, pp. 93–104. http://dx.doi.org/10.18653/v1/W19-2911.
Declerck, M., Wen, Y., Snell, J., Meade, G., Grainger, J., 2020. Unified syntax in the bilingual mind. Psychon. Bull. Rev. 27 (1), 149–154. http://dx.doi.org/10.3758/s13423-019-01666-x.
Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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). ACL, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423.
Dhar, P., Bisazza, A., 2020. Understanding cross-lingual syntactic transfer in multilingual recurrent neural networks. CoRR abs/2003.14056. arXiv:2003.14056 URL https://arxiv.org/abs/2003.14056.
Di Eugenio, B., 1996. Centering theory and the Italian pronominal system. arxiv preprint cmp-lg/9608009.
Dorr, B.J., Garman, J., Weinberg, A., 1994. From syntactic encodings to thematic roles: Building lexical entries for interlingual MT. Mach. Transl. 9 (3–4), 221–250. http://dx.doi.org/10.1007/BF00980579.
Dryer, M.S., 2005. Order of degree word and adjective. World Atlas Lang. Struct. [Online] 370–371.

Du, J., Gui, L., Xu, R., Xia, Y., Wang, X., Cambria, E., 2020. Commonsense knowledge enhanced memory network for stance classification. IEEE Intell. Syst. 35 (4), 102–109. http://dx.doi.org/10.1109/MIS.2020.2983497.

Eisner, J., 1996. Three new probabilistic models for dependency parsing: An exploration. In: 16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996. ACL, Copenhagen, Denmark, pp. 340–345, URL https://www.aclweb.org/anthology/C96-1058/.

Esuli, A., Moreo, A., Sebastiani, F., 2020. Cross-lingual sentiment quantification. IEEE Intell. Syst. 35 (3), 106–114. http://dx.doi.org/10.1109/MIS.2020.2979203.

Ferguson, C.A., 1982. Simplified registers and linguistic theory. Except. Lang. Linguist. 49, 66.

Ferrández, A., Peral, J., 2000. A computational approach to zero-pronouns in Spanish. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 166–172.

Futrell, R., Mahowald, K., Gibson, E., 2015. Quantifying word order freedom in dependency corpora. In: Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015). Uppsala University, Uppsala, Sweden, Uppsala, Sweden, pp. 91–100, URL https://www.aclweb.org/anthology/W15-2112.

Ganin, Y., Lempitsky, V.S., 2015. Unsupervised domain adaptation by backpropagation. In: Bach, F.R., Blei, D.M. (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML. In: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, Lille, France, pp. 1180–1189, URL http://proceedings.mlr.press/v37/ganin15.html.

Gass, S., 1984. A review of interlanguage syntax: Language transfer and language universals. Lang. Learn. 34 (2), 115–132. http://dx.doi.org/10.1111/j.1467-1770.1984.tb01007.x.

Gilligan, G.M., 1989. A Cross-Linguistic Approach to the Pro-Drop Parameter (Ph.D. thesis). University of Southern California.

Godard, D., 1988. La Syntaxe Des Relatives En Français. Editions du Centre national de la recherche scientifique.

Gopal, M., Jha, G.N., 2017. Zero pronouns and their resolution in sanskrit texts. In: The International Symposium on Intelligent Systems Technologies and Applications. Springer, pp. 255–267.

Gries, S.T., Kootstra, G.J., 2017. Structural priming within and across languages: A corpus-based perspective. Biling.: Lang. Cogn. 20 (2), 235–250. http://dx.doi.org/10.1017/S1366728916001085.

Grigorova, D., 2013. An algorithm for zero pronoun resolution in Bulgarian. In: Proceedings of the 14th International Conference on Computer Systems and Technologies, 2013, pp. 276–283.

Guillaume, B., de Marneffe, M.-C., Perrier, G., 2019. Conversion et améliorations de corpus du Français annotés en Universal Dependencies. Traitement Autom. Langues 60 (2), 71–95.

Hajmohammadi, M.S., Ibrahim, R., Selamat, A., Fujita, H., 2015. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. Inform. Sci. 317, 67–77. http://dx.doi.org/10.1016/j.ins.2015.04.003, URL http://www.sciencedirect.com/science/article/pii/S0020025515002650.

Hartsuiker, R.J., Beerts, S., Loncke, M., Desmet, T., Bernolet, S., 2016. Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. J. Mem. Lang. 90, 14–30. http://dx.doi.org/10.1016/j.jml.2016.03.003, URL https://www.sciencedirect.com/science/article/pii/S0749596X16000267.

Hartsuiker, R.J., Pickering, M.J., Veltkamp, E., 2004. Is syntax separate or shared between languages? Cross-linguistic syntactic priming in spanish-english bilinguals. Psychol. Sci. 15 (6), 409–414. http://dx.doi.org/10.1111/j.0956-7976.2004.00693.x.

Hauer, B., Kondrak, G., 2020. Synonymy = translational equivalence. CoRR abs/2004.13886. arXiv:2004.13886 URL https://arxiv.org/abs/2004.13886.

Hayashi, T., Fujita, H., 2020. Cluster-based zero-shot learning for multivariate data. J. Ambient Intell. Humaniz. Comput. http://dx.doi.org/10.1007/s12652-020-02268-5, URL https://doi.org/10.1007/s12652-020-02268-5.

Hewitt, J., Liang, P., 2019. Designing and interpreting probes with control tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). ACL, Hong Kong, China, pp. 2733–2743. http://dx.doi.org/10.18653/v1/D19-1275, URL https://www.aclweb.org/anthology/D19-1275.

Hewitt, J., Manning, C.D., 2019. A structural probe for finding syntax in word representations. 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). ACL, Minneapolis, Minnesota, pp. 4129–4138. http://dx.doi.org/10.18653/v1/N19-1419, URL https://www.aclweb.org/anthology/N19-1419.

Jawahar, G., Sagot, B., Seddah, D., 2019. What does BERT learn about the structure of language?. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, Florence, Italy, pp. 3651–3657. http://dx.doi.org/10.18653/v1/P19-1356, URL https://www.aclweb.org/anthology/P19-1356.

Karthikeyan, K., Zihan, W., Mayhew, S., Roth, D., 2020. Cross-lingual ability of multilingual BERT: an empirical study. In: 8th International Conference on Learning Representations, ICLR 2020. Addis Ababa, Ethiopia, URL https://openreview.net/forum?id=HJeT3yrtDr.

Kolachina, P., Ranta, A., 2019. Bootstrapping UD treebanks for delexicalized parsing. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics. Linköping University Electronic Press, Turku, Finland, pp. 15–24, URL https://www.aclweb.org/anthology/W19-6102.

Kondratyuk, D., Straka, M., 2019. 75 languages, 1 model: Parsing universal dependencies universally. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 2779–2795. http://dx.doi.org/10.18653/v1/D19-1279, URL https://www.aclweb.org/anthology/D19-1279.

Kozhevnikov, M., Titov, I., 2013. Cross-lingual transfer of semantic role labeling models. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, Sofia, Bulgaria, pp. 1190–1200, URL https://www.aclweb.org/anthology/P13-1117.

Kübler, S., McDonald, R.T., Nivre, J., 2009. Dependency Parsing. In: Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, http://dx.doi.org/10.2200/S00169ED1V01Y200901HLT002.

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., Blunsom, P., 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, Melbourne, Australia, pp. 1426–1436. http://dx.doi.org/10.18653/v1/P18-1132, URL https://www.aclweb.org/anthology/P18-1132.

Lahousse, K., Lamiroy, B., 2012. Word order in french, spanish and Italian: A grammaticalization account. Folia Linguist. 46 (2), 387–416. http://dx.doi.org/10.1515/flin.2012.014.

Lakretz, Y., Dehaene, S., King, J., 2020. What limits our capacity to process nested long-range dependencies in sentence comprehension?. Entropy 22 (4), 446. http://dx.doi.org/10.3390/e22040446.

Li, W., Zhu, L., Shi, Y., Guo, K., Cambria, E., 2020. User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models. Appl. Soft Comput. 94, 106435. http://dx.doi.org/10.1016/j.asoc.2020.106435, URL https://www.sciencedirect.com/science/article/pii/S1568494620303756.

Linzen, T., 2019. What can linguistics and deep learning contribute to each other? response to pater. Language 95 (1), e99–e108. http://dx.doi.org/10.1353/lan.2019.0001.

Linzen, T., Baroni, M., 2021. Syntactic structure from deep learning. Annu. Rev. Linguist. 7, 195–212. http://dx.doi.org/10.1146/annurev-linguistics-032020-051035.

Linzen, T., Dupoux, E., Goldberg, Y., 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Trans. Assoc. Comput. Linguist. 4, 521–535.

Liu, H., 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. Lingua 120 (6), 1567–1578. http://dx.doi.org/10.1016/j.lingua.2009.10.001, URL https://www.sciencedirect.com/science/article/pii/S0024384109002137 Contrast as an information-structural notion in grammar.

Liu, H., Xu, C., 2012. Quantitative typological analysis of romance languages. Pozn. Stud. Contemp. Linguist. 48 (4), 597–625. http://dx.doi.org/10.1515/psicl-2012-0027.

Liu, H., Xu, C., Liang, J., 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. Phys. Life Rev. 21, 171–193. http://dx.doi.org/10.1016/j.plrev.2017.03.002, URL http://www.sciencedirect.com/science/article/pii/S1571064517300532.

Loebell, H., Bock, K., 2003. Structural priming across languages. Linguistics 41 (5), 791–824. http://dx.doi.org/10.1515/ling.2003.026.

Majid, A., Jordan, F., Dunn, M., 2015. Semantic systems in closely related languages. Lang. Sci. 49, 1–18. http://dx.doi.org/10.1016/j.langsci.2014.11.002, Semantic systems in closely related languages. URL https://www.sciencedirect.com/science/article/pii/S0388000114001466.

Marchello-Nizia, C., 2006. Grammaticalisation et Changement Linguistique. De Boeck-Duculot.

Marvin, R., Linzen, T., 2019. Targeted syntactic evaluation of language models. Proc. Soc. Comput. Linguist. (SCiL) 373–374.

McCoy, R.T., Frank, R., Linzen, T., 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. Trans. Assoc. Comput. Linguist. 8, 125–140. http://dx.doi.org/10.1162/tacl_a_00304, URL https://www.aclweb.org/anthology/2020.tacl-1.9.

McWhorter, J.H., 2001. The worlds simplest grammars are creole grammars. Linguist. Typol. 5 (2–3), 125–166.

Newmeyer, F.J., 2008. Universals in syntax. Linguist. Rev. 25 (1–2), 35–82. http://dx.doi.org/10.1515/TLIR.2008.002.

Nivre, J., Fang, C.-T., 2017. Universal dependency evaluation. In: Proceedings of the NoDaLIDa 2017 Workshop on Universal Dependencies (UDW 2017). ACL, Gothenburg, Sweden, pp. 86–95, URL https://www.aclweb.org/anthology/W17-0411.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D., 2016. Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1659–1666, URL https://www.aclweb.org/anthology/L16-1262.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajic, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D., 2020a. Universal dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4034–4043.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D., 2020b. Universal dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 4034–4043, URL https://www.aclweb.org/anthology/2020.lrec-1.497.

Pamungkas, E.W., Basile, V., Patti, V., 2020. Misogyny detection in Twitter: a multilingual and cross-domain study. Inf. Process. Manage. 57 (6), 102360. http://dx.doi.org/10.1016/j.ipm.2020.102360, URL http://www.sciencedirect.com/science/article/pii/S0306457320308554.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). ACL, New Orleans, Louisiana, pp. 2227–2237. http://dx.doi.org/10.18653/v1/N18-1202, URL https://www.aclweb.org/anthology/N18-1202.

Pires, T., Schlinger, E., Garrette, D., 2019. How multilingual is multilingual BERT?. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, Florence, Italy, pp. 4996–5001. http://dx.doi.org/10.18653/v1/P19-1493, URL https://www.aclweb.org/anthology/P19-1493.

Raganato, A., Tiedemann, J., 2018. An analysis of encoder representations in transformer-based machine translation. In: Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018. ACL, Brussels, Belgium, pp. 287–297. http://dx.doi.org/10.18653/v1/w18-5431.

Ranta, A., Angelov, K., Bringert, B., 2009. Grammar development in GF. In: Proceedings of the Demonstrations Session At EACL 2009. ACL, Athens, Greece, pp. 57–60, URL https://www.aclweb.org/anthology/E09-2015.

Rasooli, M.S., Collins, M., 2017. Cross-lingual syntactic transfer with limited resources. Trans. Assoc. Comput. Linguist. 5, 279–293. http://dx.doi.org/10.1162/tacl_a_00061, URL https://www.aclweb.org/anthology/Q17-1020.

Ravishankar, V., Kulmizev, A., Abdou, M., Søgaard, A., Nivre, J., 2021. Attention can reflect syntactic structure (if you let it). In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp. 3031–3045, URL https://www.aclweb.org/anthology/2021.eacl-main.264.

Rizzi, L., 1982. Issues in Italian Syntax. Foris Publications, Dorderecht, The Netherlands.

Rizzi, L., 1986. Null objects in Italian and the theory of pro. Linguist. Inq. 17 (3), 501–557.

Rönnqvist, S., Kanerva, J., Salakoski, T., Ginter, F., 2019. Is multilingual BERT fluent in language generation?. In: Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing. Linköping University Electronic Press, Turku, Finland, pp. 29–36, URL https://www.aclweb.org/anthology/W19-6204.

Rothman, J., 2009. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. Int. J. Biling. 13 (2), 155–163.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., Villemonte de la Clergerie, E., 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages. Association for Computational Linguistics, Seattle, Washington, USA, pp. 146–182, URL https://www.aclweb.org/anthology/W13-4917.

Shin, J.-A., Christianson, K., 2009. Syntactic processing in Korean–english bilingual production: Evidence from cross-linguistic structural priming. Cognition 112 (1), 175–180. http://dx.doi.org/10.1016/j.cognition.2009.03.011, URL https://www.sciencedirect.com/science/article/pii/S0010027709000742.

Siddhant, A., Johnson, M., Tsai, H., Ari, N., Riesa, J., Bapna, A., Firat, O., Raman, K., 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. AAAI Press, New York, NY, USA, pp. 8854–8861, URL https://aaai.org/ojs/index.php/AAAI/article/view/6414.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., Manning, C., 2014. A gold standard dependency corpus for english. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, Iceland, pp. 2897–2904, URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf.

Silvestri, S., Gargiulo, F., Ciampi, M., De Pietro, G., 2020. Exploit multilingual language model at scale for ICD-10 clinical text classification. In: ISCC 2020. IEEE, Rennes, France, pp. 1–7. http://dx.doi.org/10.1109/ISCC50000.2020.9219640.

Simi, M., Bosco, C., Montemagni, S., 2014. Less is more? Towards a reduced inventory of categories for training a parser for the Italian stanford dependencies. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, Iceland, pp. 83–90.

Søgaard, A., Ruder, S., Vulić, I., 2018. On the limitations of unsupervised bilingual dictionary induction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 778–788. http://dx.doi.org/10.18653/v1/P18-1072, URL https://www.aclweb.org/anthology/P18-1072.

Solodow, J.B., 2010. Latin Alive: The Survival of Latin in English and the Romance Languages. Cambridge University Press.

Song, L., Xu, K., Zhang, Y., Chen, J., Yu, D., 2020. ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, Online, pp. 5429–5434. http://dx.doi.org/10.18653/v1/2020.acl-main.482.

Spence Green, C.S., Manning, C.D., 2009. NP subject detection in verb-initial Arabic clauses. In: Proceedings of the Third Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL3), Vol. 112, p. 123.

Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R., 2020. Anaphora and coreference resolution: A review. Inf. Fusion 59, 139–162. http://dx.doi.org/10.1016/j.inffus.2020.01.010, URL https://www.sciencedirect.com/science/article/pii/S1566253519303677.

Tenney, I., Das, D., Pavlick, E., 2019a. BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, Florence, Italy, pp. 4593–4601. http://dx.doi.org/10.18653/v1/P19-1452, URL https://www.aclweb.org/anthology/P19-1452.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., Pavlick, E., 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In: 7th International Conference on Learning Representations, ICLR 2019. New Orleans, LA, USA, URL https://openreview.net/forum?id=SJzSgnRcKX.

Thierry, G., Wu, Y.J., 2007. Brain potentials reveal unconscious translation during foreign-language comprehension. Proc. Natl. Acad. Sci. 104 (30), 12530–12535. http://dx.doi.org/10.1073/pnas.0609927104, arXiv:https://www.pnas.org/content/104/30/12530.full.pdf URL https://www.pnas.org/content/104/30/12530.

Thompson, B., Roberts, S., Lupyan, G., 2018. Quantifying semantic alignment across languages. In: Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSCi 2018), Madison, WI, USA, 2018, pp. 2551–2556.

Tsarfaty, R., Nivre, J., Andersson, E., 2012. Cross-framework evaluation for statistical parsing. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 44–54.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., Tounsi, L., 2010. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages. Association for Computational Linguistics, Los Angeles, CA, USA, pp. 1–12, URL https://www.aclweb.org/anthology/W10-1401.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, CA, USA, pp. 5998–6008, URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Vennemann, T., 1974. Topics, subjects and word order: from SXV to SVX via TVX. In: Historical Linguistics: Proceedings of the First International Congress of Historical Linguistics, Edinburgh, Scotland, pp. 339–376.

Vulić, I., Glavaš, G., Reichart, R., Korhonen, A., 2019. Do we really need fully unsupervised cross-lingual embeddings?. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 4407–4418. http://dx.doi.org/10.18653/v1/D19-1449, URL https://www.aclweb.org/anthology/D19-1449.

Wang, L., Tu, Z., Shi, S., Zhang, T., Graham, Y., Liu, Q., 2018. Translating pro-drop languages with reconstruction models. In: McIlraith, S.A., Weinberger, K.Q. (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). AAAI Press, New Orleans, Louisiana, USA, pp. 4937–4945, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16187.

Wang, L., Tu, Z., Zhang, X., Liu, S., Li, H., Way, A., Liu, Q., 2017. A novel and robust approach for pro-drop language translation. Mach. Transl. 31 (1–2), 65–87.

Warstadt, A., Singh, A., Bowman, S.R., 2019. Neural network acceptability judgments. Trans. Assoc. Comput. Linguist. 7, 625–641. http://dx.doi.org/10.1162/tacl_a_00290, URL https://www.aclweb.org/anthology/Q19-1040.

Whaley, L.J., 1996. Introduction to Typology: The Unity and Diversity of Language. SAGE publications, http://dx.doi.org/10.4135/9781452233437.

Wu, Z., Chen, Y., Kao, B., Liu, Q., 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, Online, pp. 4166–4176, URL https://www.aclweb.org/anthology/2020.acl-main.383.

Wu, S., Dredze, M., 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 833–844. http://dx.doi.org/10.18653/v1/D19-1077, URL https://www.aclweb.org/anthology/D19-1077.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C.D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., Li, J., 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. Association for Computational Linguistics, Vancouver, Canada, pp. 1–19. http://dx.doi.org/10.18653/v1/K17-3001, URL https://www.aclweb.org/anthology/K17-3001.