UNIVERSITÀ DI PISA

Dipartimento di Filologia, Letteratura e Linguistica

Corso di Laurea Magistrale in Informatica Umanistica

TESI DI LAUREA MAGISTRALE

Assessing Island Effects in Italian Transformer-based Language Models

Relatore:
Prof. Alessandro Lenci

Candidato:
Mauro Madeddu

ANNO ACCADEMICO 2021/2022

Abstract

Modern language models based on deep artificial neural networks have achieved significant progress in Natural Language Processing applications. This has spawned a line of research aimed at clarifying which linguistic phenomena and generalizations are actually learned by these models. One of the main approaches for this goal, is testing these models' sentence acceptability estimates with fine-grained targeted linguistic evaluations, based on minimal pairs that isolate a particular linguistic phenomenon.

This kind of assessment is relevant to address the open problems of the limitations that these models still have, like being significantly data-inefficient in their training, compared to humans' language acquisition and learning skills; or their still insufficient linguistic performance or generalization for some linguistic phenomena. This kind of assessment has also a broad interdisciplinary relevance since language models could be used to test theoretical linguistics hypotheses, and theoretical linguistics and psycholinguistics could in turn provide insights on how to improve these models' linguistic skills to more human-like levels.

In this work, we focus on the syntactic phenomena of island effects, and extend the Italian test suite from the psycholinguistic and experimental syntax work by Sprouse et al. (2016). Then, we evaluate on these test suites two transformer-based language models (Gpt-2 and Bert), pretrained in Italian, and compare their performance with those on humans. (..)

Contents

List of Figures

This page unintentionally left blank

Chapter 1

Introduction

1.1 Motivation

With the large progress made since 2018 in NLP benchmarks by the last generation of artificial neural network language models, which are based on a deep architecture of transformers layers and adopt a self-supervised learning approach, it has become increasingly important to understand what these model actually learn about human language (Hewitt and Manning, 2019; Manning et al., 2020; Trotta et al., 2021).
Investigating this serves multiple purposes, like providing insights for building better models, but it is also of interest for addressing research questions in linguistics (Hewitt and Manning, 2019).

(TODO: virgolettati da sintetizzare di seguito)

"Linguistic competence of neural language models (LMs) has emerged as one of the core sub-fields in NLP. " (Cherniavskii et al. 2022 Acceptability Judgements via Examining the Topology of Attention Maps)

"more fine-grained evaluation tools may accelerate work on general-purpose neural network modules for sentence understanding. "
"studying the linguistic competence of ANNs bears on foundational questions in linguistics about the learnability of grammar."

(black boxes and explainability) : "the nature of the representations learned by these models is not properly understood." (Wilcox et al., 2018)

"In addition to the insight these results provide about neural NLP systems, they also bear on questions central to cognitive science and linguistics, putting lower bounds on what syntactic knowledge can be acquired from string input alone" (Hu et al., 2020)

1.2 Research on linguistic tests on language models

(TODO)

1.3 Island Effects and their assessment

1.3.1 Overview on island effects

(TODO: sintetizzare i virgolettati di seguito)

"filler–gap dependencies are subject to numerous complex island constraints: Ross (1968) identified five syntactic positions in which gaps are illicit, dubbing them syntactic islands." (Wilcox et al., 2018)

Examples: (..)

"open question whether these “island constraints” are true grammatical constraints, or whether they are effects of processing difficulty or discourse-structural factors (Ambridge and Goldberg, 2008; Hofmeister and Sag, 2010; Sprouse and Hornstein, 2014)" (Wilcox et al., 2018)

Argument from the poverty of the stimulus (APS): "Because of their complexity and ubiquity, these dependencies have figured prominently in arguments that natural language would be unlearnable by children without a great deal of innate knowledge (Phillips, 2013) (cf. Pearl and Sprouse, 2013; Ellefson and Christiansen, 2000)" (Wilcox et al., 2018)

"The influential argument from the poverty of the stimulus (APS) .. has been subject to much criticism (Pullum and Scholz, 2002)" Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. The Linguistic Review, 18(1-2):9–50.

Wh-Island Constraint definition: "A gap cannot appear inside doubly nested clauses headed by wh-complementizers. This phenomenon is called the Wh-Island Constraint (WHC). " (Wilcox et al., 2018)

1.3.2 Island effects assessment in language models

(TODO)

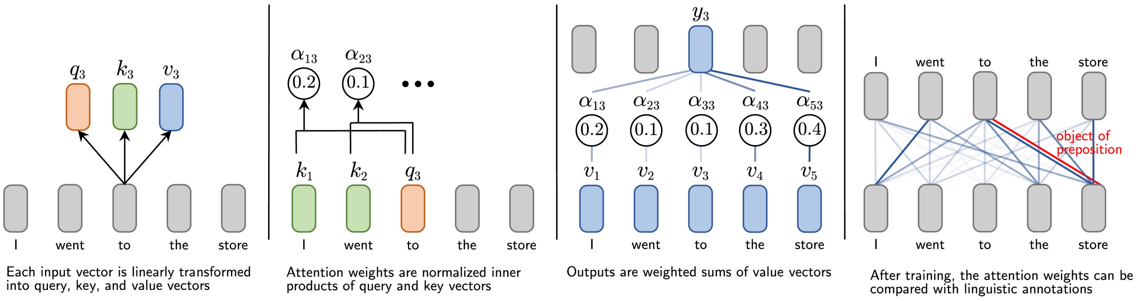
1.4 Transformer-based language models

1.4.1 Transformers and the Attention Mechanism

Transformers are "a neural network architecture, without any recurrent connections (22), which takes a sequence of words (or other symbols) as input and produces a contextualized vector representation of each word as its out-put (Fig. 3). It contains many millions of trainable parameters in a number of layers, typically requiring massive amounts of data and computation to train. This makes Transformers difficult to train, but also highly expressive models that can out-perform other contemporary neural networks when properly optimized" (Manning et al., 2020)

"The key mechanism by which Transformers contextualize representations is multi-headed attention (see Fig. 5)." (Manning et al., 2020) "attention is a kind of **generalized dot product**" which models all pairwise interactions between words.

"Attention(23) dynamically assigns a weight to every pair of words in the sequence, indicating how much the model should “pay attention to” the first word when computing the representation of the second one." (Manning et al., 2020) "Transformers use multiple attention heads in parallel, where each head can potentially capture a completely different word–word relation. Transformers aggregate the information from each head to produce a single output vector representation for each word in the sequence." (Manning et al., 2020)



Each input vector is linearly transformed into query, key, and value vectors

Attention weights are normalized inner products of query and key vectors

Outputs are weighted sums of value vectors

After training, the attention weights can be compared with linguistic annotations

Figure 1.1: Overview of the attention mechanism, taken from Manning et al. (2020) (TODO: request permission, pnas licence is CC ND)

..

The lack of recurrence in the transformers architecture "enables greater within-training-example parallelization, at the cost of quadratic complexity in the input sequence length" (Liu et al., 2018)

the terminology used to describe the Transformer: "the attention is a function of a query (Q) and set of key (K) and value (V) pairs" (Liu et al., 2018)

Training objective function, cross-entropy loss: "During training, the probability assigned to the correct word is used to calculate the cross-entropy loss for each item in the sequence. As with RNNs, the loss for a training sequence is the average cross-entropy loss over the entire sequence." (Jurasfky 3rd ed. ch.9)(figure 9.21)

transformer blocks "Each provides a multi-headed self-attention unit over all input words, allowing it to capture multiple dependencies between words, while avoiding the need for recurrence. With no need to process a sentence in sequence, the model parallelizes more efficiently, and scales in a way that RNNs cannot" (Lau et al., 2020)

1.4.2 Traditional unidirectional language models (Gpt2)

Conventional language models, like LSTM, TDLM, and GPT2, are unidirectional, they predict the probability of a token using only past tokens "the input [to GPT2] is a sequence of previously seen words, which are then mapped to embeddings (along with their positions) and fed to multiple layers of "transformer blocks" before the target word is predicted" ""Much of its power resides in these transformer blocks" (Lau et al., 2020)

The unidirectionality of conventional LMs allows to estimate log probabilities for a sentence W via the chain rule (Salazar et al., 2020) $logP_{LM}(W) = \sum_{t=1}^{|W|} logP_{LM}(w_t|W_{<t}$"

"Given a unidirectional language model, we can infer the probability of a sentence by multiplying the estimated probabilities of each token using previously seen (left) words as context (Bengio et al., 2003):" (Lau et al., 2020)

..

1.4.3 Bidirectional language models (BERT)

"BERT (Devlin et al., 2019) and its improvements to natural language understanding have spurred a rapid succession of contextual language representations (Yang et al., 2019; Liu et al., 2019; inter alia) which use larger datasets and more involved training schemes. Their success is attributed to their use of bidirectional context, often via their masked language model (MLM) objectives." Salazar et al. (2020)

"The self-supervision task used to train BERT is the masked language-modeling or cloze task, where one is given a text in which some of the original words have been replaced with a special mask symbol. The goal is to predict, for each masked position, the original word that appeared in the text (Fig. 3).To perform well on this task, the model needs to leverage the surrounding context to infer what that word could be." (Manning et al., 2020)

"BERT's MLM objective can be viewed as stochastic maximum pseudolikelihood estimation (MPLE) Wang and Cho (2019)(Besag, 1975) on a training set .. " (Salazar et al., 2020) "In this way, MLMs learn an underlying joint distribution whose conditional distributions $w_t|W_{\setminus t}$ are modeled by masking at position t." (Salazar et al., 2020)

Chapter 2

Related work

The previous works closest to ours are those by Wilcox et al. (2018); Hu et al. (2020); Sprouse et al. (2016), which use a factorial test design and test on island effects, on neural language models for Wilcox et al. (2018); Hu et al. (2020), and on human subjects for Sprouse et al. (2016)

Hu et al. (2020) uses a factorial design with minimal pairs and obtains a percentage accuracy score. An item is considered to have been scored by the models when multiple conditions are met (eg. the unacceptable sentence is scored lower than the others, and a factorial effect measurement is greater than zero). However, the phenomena tested by Hu et al. (2020) can be formulated in the stringent way of minimal pairs differing for just one word; which is not well-suited for island effects.

Wilcox et al. (2018); Sprouse et al. (2016) instead obtain a measure of statistical significance and a confidence interval. In the case of Sprouse, however, the format of sentences employed is not suitable for scoring language models, which require that all the sentences in an item are minimally different in terms of lexical content, and ideally should differ in just one word. Wilcox et al. (2018) circumvents this issue by scoring separately items with different construct (i.e. one with and one without an island structure) which are lexically very similar but differ more than a minimal pair. For both items, they measure the effect of the filler-gap dependency phenomena, whose effect is expected to drop in the presence of an island construct, which blocks it. If the drop in the effect is statistically significant, they can conclude that the model has "learned" the island constraint.

Both Wilcox et al. (2018); Hu et al. (2020) tested only conventional left-to-right unidirectional language models, none tested on bidirectional language models like BERT.

2.1 Approaches for targeted linguistic evaluation of language models

Linguists concerned with acceptability judgements, while "Computational language processing has traditionally been more concerned with *likelihood*—the probability of a sentence being produced or encountered" Lau et al. (2020) "The question of whether and how these properties are related is a fundamental one" (Lau et al., 2020)

"Acceptability is closely related to the concept of grammaticality. The latter is a theoretical construction corresponding to syntactic wellformedness, and it is typically interpreted as a binary property (i.e., a sentence is either grammatical or ungrammatical). Acceptability, on the other hand, includes syntactic, semantic, pragmatic, and non-linguistic factors, such as sentence length. It is gradient, rather than binary, in nature (Denison, 2004; Sorace and Keller, 2005; Sprouse, 2007)" (Lau et al., 2020)

"Overview of the approach" "Grammaticality and LM probability" (Marvin and Linzen, 2018)

"How should grammaticality be captured in the probability distribution defined by an LM? The most extreme position would be that a language model should assign a probability of zero to ungrammatical sentences. For most applications, some degree of error tolerance is desirable, and it is not practical to assign a sentence a probability of exactly zero.1 " "1Nor is it possible to have a threshold episilon such that all grammatical sentences have probability higher than episilon and all ungrammatical sentences have probability lower than episilon, for the simple reason that there is an infinite number of grammatical sentences (Lau et al., 2017)." (Marvin and Linzen, 2018)

The Corpus of Linguistic Acceptability (CoLA) is the predecessor of Blimp minimal pairs approach Salazar et al. (2020), it has a training set and asks to label sentences as “acceptable” or not in isolation, with an absolute score Warstadt et al. (2019)" Salazar et al. (2020).

In minimal pairs approaches, on the other hand, there is a relative comparison bewtween two sentences, and the model is considered to have given the right answer if it gives an higher score to the acceptable sentence than the unacceptable (or more marginal) one.

Another fundamental difference is that minimal pairs scoring is usually done with the pretrained models out of the box, while acceptability judgments datasets like CoLA require the models to be finetuned on an additional dataset on an acceptability detection task.

"This task is often employed to evaluate language models because the outputted probabilities for a pair of minimally different sentences are directly comparable, while the output for a single sentence cannot be taken as a measure of acceptability without some kind of normalization (Lau et al., 2016)"

"Accordingly, CoLA, but not datasets based solely on preferences between minimal pairs, may be used to evaluate models' ability to make judgments that align with both native speaker judgments and the predictions of generative theories."

2.1.1 Acceptability judgments (absolute scores)

"Acceptability judgments are the main form of behavioral data used in generative linguistics to measure human linguistic competence (Chomsky, 1965; Schutze, 1996)." (Warstadt et al., 2020)

..

2.1.2 Minimal pairs (relative scores)

"the LM probability of a sentence can only serve as a proxy for acceptability if **confounding factors impacting a sentence's probability** such as **length** and **lexical content** are **controlled** for. It is with these considerations in mind that we design BLiMP" (Warstadt et al., 2020)

"Since most minimal pairs [in BLiMP] only differ by a single word, the effect of length on log probabilities and PLLs (discussed in Section 4.3) is mitigated" (Salazar et al., 2020)

2.1.3 Factorial approaches

Factorial experimental setup example: (..)

Hu et al. (2020) use these controlled tests to "describe and test for human-like syntactic knowledge in language models" The testing paradigm presented by Hu et al. (2020) "**compare critical sentence regions instead of full-sentence probabilities**, and employ a 2 × 2 paradigm with a strict, multi-fold success criterion inspired by psycholinguistics methodology" (Hu et al., 2020) This allows them "to factor out as many confounds as possible, such as the lexical frequency of individual tokens and low-level n-gram statistics." (Hu et al., 2020)

"Each test suite contains a number of ITEMS (typically between 20 and 30), and each item appears in several CONDITIONS: across conditions, a given item will differ only according to a controlled manipulation designed to target a particular feature of grammatical knowledge. " (Hu et al., 2020) "Each test suite contains at least one PREDICTION, which specifies inequalities between **surprisal** values at **pairs of regions/conditions** that should hold if a model has learned the appropriate syntactic generalization. We expect language models which have learned the appropriate syntactic generalizations from their input to **satisfy these inequalities** without further fine-tuning. We compute **accuracy** on a test suite as the proportion of items for which the model's behavior conforms to the prediction " (Hu et al., 2020)

In Hu et al. (2020) "a model's accuracy for a test suite is computed as the percentage of the test suite's items for which it satisfies the criterion."

In Hu et al. (2020), "Test suites are written using a standard format that allows for flexible predictions which more **closely resemble those used in psycholinguistic studies**, specifically allowing for predictions about **interactions among multiple testing conditions**."

"In this paper, we look for cases where the surprisal associated with an an unusual construction—such as a gap—is ameliorated by the presence of a licensor, such as a wh-word." (Wilcox et al., 2018) "If the models learn that syntactic gaps require licensing, then sentences with licensors should exhibit lower surprisal than minimally different pairs that lack a proper licensor." (Wilcox et al., 2018)

"We test whether the LSTM language models have learned filler–gap dependencies by looking for a 2x2 interaction between the presence of a gap and the presence of a wh-licensor"

"We use experimental items where the gap is located in an obligatory argument position, e.g. in subject position or as the direct object of a transitive verb, **as judged by the authors**." (Wilcox et al., 2018) "The phrase with the gap is embedded inside a complement clause. We chose this paradigm over bare wh-questions because **it eliminates do-support and tense manipulation of the main verb**, resulting in **higher similarity across conditions**" (Wilcox et al., 2018)

"In the following experiments, we examine whether RNN language models have learned constraints on filler–gap dependencies by comparing the wh-licensing interaction in non-islands to that within islands. The strongest evidence for an island constraint would be if the wh-licensing interaction goes to zero for a gap in island position, implying that, in the distribution over strings implied by the network, the appearance of a whlicensor is totally unrelated to the appearance of a gap in the island position" (Wilcox et al., 2018)

"More generally, we can look for a weakened wh-licensing interaction for island vs. non-island positions, which would mean that the network believes a relationship between the wh-licensor and the island gap is less likely" (Wilcox et al., 2018) "A positive but nonzero wh-licensing interaction would be in line with human acceptability judgments, which do not always categorically rule out gaps in island positions (Ambridge and Goldberg, 2008), and with human online processing experiments, which have shown that gap expectation is attenuated during processing of areas where gaps cannot occur licitly, but does not always disappear entirely (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006). Therefore, in this section we take a significant reduction in the island relative to the non-island case to constitute evidence that the model has 'learned' the constraint." (Wilcox et al., 2018)

2.2 Linguistic evaluation measures for language models

"To accommodate sentence length and lexical frequency we experiment with several simple normalization methods, converting probabilities to acceptability measures (Section 3.2)." (Lau et al., 2020)

2.2.1 Surprisal

In probabilistic language models, garden-path disambiguation effects "are well captured by word negative log probabilities, or **SURPRISALS** (Hale, 2001): $S(w|C) = -log2p(w|C)$, " (Hu et al., 2020) Surprisals "are independently well-established to predict human incremental processing difficulty over several orders of magnitude in word probability (Smith and Levy, 2013). " (Hu et al., 2020)

"evaluate MLMs out of the box via their pseudo-log-likelihood scores (PLLs)," (Salazar et al., 2020) called ..LP by (Lau et al., 2020) and .. by prev study .. (2016) " pseudo-loglikelihood scores (PLLs) from MLMs (Wang and Cho, 2019), " (Salazar et al., 2020)

"PLL's unsupervised expression of linguistic acceptability without a left-to-right bias, greatly improving on scores from GPT-2 (+10 points on island effects, NPI licensing in BLiMP) " (Salazar et al., 2020)

"To score a sentence, one creates copies with each token masked out. The log probability for each missing token is summed over copies to give the pseudo-log-likelihood score (PLL)." (Figure from Salazar et al. (2020)) "given by summing the conditional log probabilities log PMLM(wt j Wnt) of each sentence token (Shin et al., 2019). These are induced in BERT by replacing wt with [MASK]" (Salazar et al., 2020)

PLL's summands are conditional probabilities (Salazar et al., 2020) "Log probabilities model the joint distribution; PLL does so as well, albeit implicitly (Appendix B)" (Salazar et al., 2020)

" domain shifts can be visibly observed from the positionwise scores log PMLM(wt j Wnt)" (Salazar et al., 2020)

" by learning g(W) first. They argue g expresses **fluency**; fixing$g$ early allows f(W ; X) to focus its capacity on **adequacy** in encoding the source, and thus specializing the two models. With this perspective in mind, we compare log PLM and PLL as candidates for log g." (Salazar et al., 2020) "In this work we interpret fluency as linguistic acceptability .. informally, the syntactic and semantic validity of a sentence according to human judgments (Schutze ¨ , 1996)" (Salazar et al., 2020) "Its graded form is well-proxied by neural language model scores (log PLM) once length and lexical frequency are accounted for (Lau et al., 2017)." (Salazar et al., 2020)

2.2.2 Issues with perplexity

Perplexity: "perplexity on large benchmark datasets like WikiText-103 (Merity et al., 2016) has remained the primary performance metric, which cannot give detailed insight into these models' knowledge of grammar." (Warstadt et al., 2020) "the most widespread currency of evaluation for language models is perplexity-how well, on average, a model predicts a word in its context" (Hu et al., 2020)

"In previous work, perplexity and syntactic judgment accuracy have been found to be partly dissociable (Kuncoro et al., 2018; Tran et al., 2018)" (Marvin and Linzen, 2018) Hu et al. (2020) found a dissociation between perplexity and syntactic generalization performance.

"The quality of the syntactic predictions made by the LM is arguably particularly difficult to measure using perplexity: since most sentences are grammatically simple and most words can be predicted from their local context, perplexity rewards LMs primarily for collocational and semantic predictions." (Marvin and Linzen, 2018)

"a broad-coverage metric such as perplexity may not be ideal for assessing human-like syntactic knowledge for a variety of reasons. In principle, a sentence can appear with vanishingly low probability but still be grammatically wellformed, such as *Colorless green ideas sleep furiously* (Chomsky, 1957). " (Hu et al., 2020) "While perplexity remains an integral part of language model evaluation, fine-grained linguistic assessment can provide both more challenging and more interpretable tests to evaluate neural models." (Hu et al., 2020)

2.2.3 Pseudo-log-likelihood scores (PLLs)

The Formula for the sentence acceptability estimates $LP, PenLP$ from Lau et al. (2020) are: (..)

"pseudo-log-likelihood scores (PLLs)" "are computed by masking tokens one by one" and summing up the resulting log probabilities of each masked token (Salazar et al., 2020)

"Unlike log probabilities, PLL's summands are more uniform across an utterance's length (no left-toright bias), helping **differentiate fluency from likeliness**" (Salazar et al., 2020)

Lau et al. (2020) formula for BERT sentence acceptability: "It is important to note, however, that sentence probability computed this way is not a true probability value: These probabilities do not sum to 1.0 over all sentences. Equation (1), in contrast, does guarantee true probabilities" (Lau et al., 2020) On BERT formula: "Intuitively, the sentence probability computed with this bidirectional formulation is a measure of the model's confidence in the likelihood of the sentence"(Lau et al., 2020) "To compute the true probability, Wang and Cho (2019) show that we need to sum the pre-softmax weights for each token to score a sentence, and then divide the score by the total score of all sentences. As it is impractical to compute the total score of all sentences (an infinite set), the true sentence probabilities for these bidirectional models are intractable. We use our non-normalized confidence scores as stand-ins for these probabilities."(Lau et al., 2020)

"Sentence probability (estimated either using unidirectional or bidirectional context) is affected by its length (e.g., longer sentences have lower probabilities), and word frequency (e.g., the cat is big vs. the yak is big)" (Lau et al., 2020)

We calculated the LP and PenLP measures taking as a basis the models' outputs after the softmax activation function applied to the last layer.

The formulas for LP and PenLP (taken from (Lau et al., 2020))are:

$$LP = \log P(s)$$

$$PenLP = \frac{LP}{((5 + |s|)\big/(5 + 1))^{\alpha}}$$

Where $P(s)$ is the probability of the sentence s. The PenLP divides the LP estimate by a penalty term which depends on the sentence lenght $|s|$, measured in number of tokens. Following Lau et al. (2020), we use $\alpha = 0.8$.

For BERT-like models, we use the estimate from Lau et al. (2020) of the log probability LP of a sentence s. It is estimated by masking each word in s, calculating the probability of the prediction of the masked word, and summing the log of all these probabilities:

..

Numerical properties of PLL fixing |W|, flat graph for Bert, descending curve for Gpt "given fixed jWj one expects -log PMLM(wt j Wnt) to be in the same range for all t. Meanwhile -log PLM(wt j W<t) decreases as t ! jWj, the rate of which was studied in recurrent language models (Takahashi and Tanaka-Ishii, 2018)." Salazar et al. (2020) "the outsized cost of the unconditional first unigram in Figure 3. " "All MLMs spike at the final token of an utterance, before our appended period ".". Terminal words are difficult to predict in general, " "Otherwise, the averaged cross-entropies are flat." Salazar et al. (2020) (TODO: insert the 2 figures from salazar et al.)

"averaged cross-entropies are flat.": "This, plus our success on BLiMP, suggest positionwise scores as a way of detecting "disfluencies" (at least, those in the form of domain mismatches) by observing spikes in cross-entropy; " Salazar et al. (2020) "with log PLM, spikes are confounded by the curve in Figure 3." Salazar et al. (2020) "In Appendix C, we plot sentence-level PLLs versus jWj and observe linearity as jWj ! 1, with spikes from the last word and lowercase first word smoothing out." "This behavior motivates our choice of $\alpha = 1 : 0$ when applying the Google NMT-style length penalty (Wu et al., 2016) to PLLs, which corresponds to the asymptoticallylinear LPMLM = (5 + jWj)=(5 + 1). " Salazar et al. (2020)

the Google NMT-style length penalty (Wu et al., 2016) Salazar et al. (2020)

2.3 Test suites

2.3.1 CoLA

On test suite development: "investigates acceptability judgments in real textual contexts, " (Lau et al., 2020) "naturalistic approach vs a constructed dataset" "syntactically challenging examples are sparsely represented in a corpus" "naturally occurring sentences are difficult to control for confounds" (Marvin and Linzen, 2018)

Corpora of sentences and their grammaticality. "The most recent and comprehensive corpus is CoLA (Warstadt et al., 2019b), containing 10k sentences covering a wide variety of linguistic phenomena provided as examples in linguistics papers and books."(Warstadt et al., 2020) CoLA is included in the GLUE benchmark (Wang et al., 2018)

..

2.3.2 BLiMP

"the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), a challenge set of 67k pairs which isolate contrasts in syntax, morphology, and semantics" (Salazar et al., 2020) "BLiMP provides an unsupervised setting: language models are evaluated on how often they give the acceptable sentence a higher (i.e., less negative) score." (Salazar et al., 2020)

2.3.3 Other (Hu et al, Wilcox et al, ..)

..

2.4 Review of results from previous related work

"Current models like BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) now learn to give acceptability judgments that approach or even exceed individual human agreement with CoLA." (Warstadt et al., 2020)

"We show that PLLs outperform scores from autoregressive language models like GPT-2 in a variety of tasks. " (Salazar et al., 2020)

"Our circuit-level analysis reveals consistent failure on Licensing but inconsistent behavior on other circuits, suggesting that different syntactic circuits make use of different underlying processing capacities." (Hu et al., 2020)

Warstadt et al. (2020) (which did not test BERT): "current neural LMs appear to acquire robust knowledge of morphological agreement and some syntactic phenomena such as ellipsis and control/raising. They show weaker evidence of knowledge about argument structure, negative polarity item licensing, and the semantic properties of quantifiers. **All models perform at or near chance on extraction islands**. Overall, every model we evaluate falls short of human performance by a wide margin" "ISLANDS are the hardest phenomenon by a wide margin. Only GPT-2 performs well above chance, and it remains 14 points below humans." Warstadt et al. (2020)

"We find, in accordance with Wilcox et al. (2018), that LMs do represent long-distance wh-dependencies, but we also conclude that their representations differ fundamentally from humans'." Warstadt et al. (2020) "evidence that increased dependency length and the presence of agreement attractors of the kind investigated by Linzen et al. (2016) and Gulordava et al. (2019) reduce performance on agreement phenomena." (on LSTM) Warstadt et al. (2020) "Although some models approach human performance in ordinary filler-gap dependencies, they are exceptionally poor at identifying island violations overall" Warstadt et al. (2020)

Warstadt et al. (2020) : "strong conclusions about how these models represent wh-dependencies are not possible using the forced-choice task compatible with BLiMP, and a **complete assessment of syntactic islands** is **best addressed using a factorial** design that manipulates both the presence of an island and an attempt to extract from it, as in Kush et al. (2018) or Wilcox et al. (2018)"

"These results demonstrate that neither [RNN] model has learned the subject constraint, categorizing PPs as either licit extraction domains in all positions (the Google model) or treating them like islands (the Gulordava model)" (Wilcox et al., 2018)

"Both [RNN] models failed to correctly generalize island constraints in two conditions: The Google model failed to learn that-headed Complex-NP Islands, the Gulordava model to learn Wh-Islands, and both failed to learn Subject Islands."(Wilcox et al., 2018)

"In other recent work, Chowdhury and Zamparelli (2018) tested the ability of neural networks to separate grammatical from ungrammatical extractions using similar metrics to ours, finding that their neural networks do not represent the unboundedness of filler–gap dependencies nor certain strong island constraints." "We believe the difference between our results and theirs is **due to experimental design**: They choose to measure the probability of the question mark punctuation as a proxy for the RNNs gap expectation, and use **sentence schemata** instead of hand-engineered experimental items."

"While Chowdhury and Zamparelli (2018) conclude that the networks are not learning island-like constraints, but rather displaying sensitivity to syntactic complexity plus order, we demonstrate island-like effects where both the island and the non-island item are equally complex (in e.g. wh-islands). " (Wilcox et al., 2018)

"While inconsistent with the formal linguistic literature on filler–gap dependencies, the negative values of all but one of the correlations are consistent with known effects in human sentence processing, where increasing distance between fillers and gaps usually causes processing slowdown (Grodner and Gibson, 2005; Bartek et al., 2011)." (Wilcox et al., 2018)

"Our work shows these dependencies and their constraints can be learned to some extent by a generic sequence model with no obvious inductive bias for hierarchical structures. This is evidence against the idea that such an inductive bias is necessary for language learning, although the amount of data these models are trained on is much larger than the typical input to a child learner." (Wilcox et al., 2018)

Salazar et al. (2020) found that, on the Blimp testset, Roberta large learns island effects close to a human level (accuracy scores are 83.4% from Roberta and 84.9% by humans). But as pointed out in the original Blimp paper (Warstadt et al., 2020), island effects phenomena are probably better tested with a factorial design like (Wilcox et al., 2018)

the improvement of RoBerta over Gpt on island effect, approaching human levels "suggests that the difficulty of these BLiMP categories was due to PLM decomposing autoregressively, and not intrinsic to unsupervised language model training, as the original results may suggest (Warstadt et al., 2020)." (Salazar et al., 2020)

"For some intuition, we include examples in Table 8. In the subject-verb agreement example, BERT sees The pamphlets and resembled those photographs when scoring have vs. has, whereas GPT-2 only sees The pamphlets, which may not be enough to counter the misleading adjacent entity Winston Churchill at scoring time" (Salazar et al., 2020) "Who does Amanda find while thinking about Lucille? Who does Amanda find Lucille while thinking about?" "GPT-2 and BERT both promote fluency, but GPT-2's left-to-right biased scores appear to cause it to **overweigh common word sequences** at the expense of adequacy" (Salazar et al., 2020)

2.5 Our research questions and contributions

(TODO)

Chapter 3

Experimental setup

3.1 Tested models

BERT (https://huggingface.co/dbmdz/bert-base-italian-xxl-cased): **81GB** (13 billion tokens) of training data and from Wikipedia, OPUS and OSCAR corpora. Model 424 MB.

GePpeTto (LorenzoDeMattei/GePpeTto): **13.8GB** of training data from Wikipedia and ItWac corpus. The model's size corresponds to GPT-2 small, with 12 layers and 117M parameters. Vocab size 30k. 620k training steps.

GilBERTo (https://huggingface.co/idb-ita/gilberto-uncased-from-camembert): Trained on **71GB** of Italian text (11.2 billion tokens) from the OSCAR corpus. Model size 420 MB

3.2 Methodology

3.3 Factorial test design

3.3.1 Introduction

Here are two sample items for the adjunct island complex NP island phenomena from the test suite we developed for this thesis:

Example 3.3.1. Adjunct islands

a. Short-NonIsland:
Chi dice che l'autore avrebbe inviato il libro all'editore?
('Who says that the author had sent the book to the publisher?')

b. Long-NonIsland:
Che cosa dici che l'autore avrebbe inviato all'editore?
('What do you say that the author had sent to the publisher?')

c. Short-Island:
Chi ha stampato l'illustrazione dopo che l'autore ha inviato il libro all'editore?
('Who printed the illustration after the author sent the book to the publisher?')

d. Long-Island:
Che cosa il disegnatore ha stampato l'illustrazione dopo che l'autore ha inviato all'editore?
('What did the designer printe the illustration after the author sent to the publisher?')

Example 3.3.2. Complex NP islands

a. Short-NonIsland:
Chi ha smentito che l'agenzia avrebbe diffuso il sondaggio?
('Who denied that the agency had released the poll?')

b. Long-NonIsland:
Cosa hai smentito che l'agenzia avrebbe diffuso?
('What have you denied that the agency had released?')

c. Short-Island:
Chi ha smentito la voce che l'agenzia avrebbe diffuso il sondaggio?
('Who denied the rumor that the agency had released the poll?')

d. Long-Island:
Cosa hai smentito la voce che l'agenzia avrebbe diffuso?
('What have you denied the rumor that the agency had released?')

Example 3.3.3. Whether islands

a. Short-NonIsland:
Chi pensa che io abbia riscosso il pagamento?

b. Long-NonIsland:
Cosa pensi che io abbia riscosso?

c. Short-Island:
Chi si domanda se io abbia riscosso il pagamento?

d. Long-Island:
Cosa ti domandi se io abbia riscosso?

As in Sprouse et al. (2016), this factorial design is aimed at isolating two factors that could impact the acceptability of a sentence: the effect of having a long-distance dependency (e.g. a wh-dependency), and the effect of having a complex syntactic structure like a syntactic island. Each item as item 3.3.1 has four sentences given by the combination of these two factors.

In item 3.3.1, the Short-NonIsland sentence has a short distance dependency, because the arguments of the main clause verb "dice" are next to it. The Long-NonIsland sentence, on the other hand, has a long-distance dependency, because the interrogative "Che cosa" depends to the verb of the subordinate clause "avrebbe inviato", as a direct object "gap". ..

In the present thesis, we focus on four phenomena of island effect structures: whether islands, complex np islands, subject islands, and adjunct islands, all based on wh-dependencies (we leave rc-dependencies like those treated in Sprouse et al. (2016), for future work.)

Using factorial sentences as minimal pairs

In the present thesis, we also use the four sentences of each item of the factorial test design, to draw three minimal pairs, comparing the three acceptable sentences (long and short non-island, and the short-island types) with the unacceptable one (the long-nonisland sentence type). We then score these minimal pairs for accuracy, considering that a model scores accurately a pair of sentences if it gives an higher acceptability score to the acceptable sentence rather than the unacceptable one.

..

3.3.2 Test suites

For the present thesis, we develop a new test suite that follows the same paradigm form as the one in Sprouse et al. (2016), but with an increased item number of 50 (from the original 8). As in the original test suite in Sprouse et al. (2016), there are four island phenomena (whether islands, complex np islands, subject islands, and adjunct islands). Each phenomena is exemplified in 50 items, which in turn are composed by four sentences as in item 3.3.1, covering the combinations of two factors: presence or absence of an island structure, and a short or long-distance dependency.

Scores normalization

To be more directly comparable with the results in Sprouse et al. (2016), the scores (LP or PenLP) were then discretized into a 7-point likert scale (using bins rather than quantiles) and normalized into z-scores.

Since for humans the likert scale discretization and the z-score normalization were done normalizing all the score of each individual separately, we considered a model (like a Gpt-2 instance) with a particular sentence acceptability approximation (LP or PenLP) as the equivalent of a human subject, normalizing in the same scale all its scores across the four wh-dependency island effects phenomena of a particular test suite.

Chapter 4

Results

4.1 Accuracy results for island effects minimal pairs in Italian

In Table 4.1 we see the accuracy scores of several Gpt and BERT-based models on an Italian island effects test suite developed for the present thesis. The minimal pairs are drawn from a factorial test design whose item are like item 3.3.1.
Each acceptable sentence of an item like item 3.3.1 is compared with the unacceptable sentence, and the score is considered accurate if the acceptable sentence receives an higher score than the unacceptable one.

For instance, on the first row of results, we see the scores for the adjunct island phenomenon (with wh-dependencies), where the acceptable Short-NonIsland sentence gets scored higher than the unacceptable sentence (Long-Island) 96 % of the times by the Gpt-2 model with the LP sentence acceptability estimate.

All models seem to struggle with the Short-Island sentence type of the Subject islands phenomenon, with the highest score being 68%. (TODO: see which of the 50 sentences are scored higher/lower, and why).

On average, the Gpt-2 model with the PenLP scoring measure performs best, with a 86.1% mean accuracy.

Although it's between different languages (English and Italian), the scores on item 3.3.1 seem to be higher than the ones on the BLiMP benchmark in Table 4.3.

..

		Gpt2 (it)		**BERT (it)**				**GilBERTo (it)**			
Phenomenon	**Sentence form**	LP	Pen LP	LP	Pen LP	LP-L	Pen LP-L	LP	Pen LP	LP-L	Pen LP-L
Wh-adjunct	Short-N.I.	**96**	92	94	90	**96**	**96**	86	70	86	86
	Long-N.I.	**98**	86	68	42	60	60	64	34	4	2
	Short-I.S.	96	98	**100**	98	**100**	**100**	94	94	84	88
Wh-complex np	Short-N.I.	90	92	**100**	**100**	96	96	74	76	88	88
	Long-N.I.	**100**	42	96	92	70	64	62	28	32	28
	Short-I.S.	38	88	**100**	**100**	96	96	46	82	88	88
Wh-subject	Short-N.I.	**98**	90	26	6	28	28	70	46	28	22
	Long-N.I.	**100**	98	86	56	78	74	76	50	24	20
	Short-I.S.	40	56	62	60	**68**	**68**	52	56	**68**	**68**
Wh-whether	Short-N.I.	91.5	94.9	94	90	**96**	**96**	91.5	94.9	89.8	89.8
	Long-N.I.	**100**	**100**	66	40	60	58	**100**	98.3	78	78
	Short-I.S.	59.3	96.6	**100**	98	**100**	**100**	37.3	69.5	93.2	93
Average		83.9%	**86.1%**	85%	78.4%	81.5%	80.7%	71.1%	66.6%	63.6%	62.6%

Table 4.1: Accuracy percentages for Gpt-2 and BERT Italian models, on a test suite of 50 items per phenomenon developed for the present thesis. The Gpt2-it model is LorenzoDeMattei/GePpeTto. The BERT-it model is dbmdz/bert-base-italian-xxl-cased. The GilBERTo-it model (an Italian RoBERTa variant) is idb-ita/gilberto-uncased-from-camembert.

		Gpt2 (it)		**BERT (it)**				**GilBERTo (it)**			
Phenomenon	**Sentence form**	LP	Pen LP	LP	Pen LP	LP-L	Pen LP-L	LP	Pen LP	LP-L	Pen LP-L
Wh-adjunct	Short-N.I.	25	75	**87.5**	**87.5**	75	75	62.5	**87.5**	**87.5**	**87.5**
	Long-N.I.	**75**	**75**	**75**	**75**	62.5	62.5	50	62.5	**75**	**75**
	Short-I.S.	25	50	**100**	**100**	**100**	**100**	12.5	50	50	62.5
Wh-complex np	Short-N.I.	**100**	**100**	**100**	**100**	**100**	**100**	62.5	50	75	75
	Long-N.I.	**100**	50	**100**	**100**	**100**	87.5	87.5	37.5	12.5	12.5
	Short-I.S.	37.5	75	**100**	**100**	**100**	**100**	62.5	87.5	87.5	87.5
Wh-subject	Short-N.I.	**100**	87.5	62.5	12.5	37.5	37.5	75	25	12.5	0
	Long-N.I.	**100**	**100**	87.5	87.5	87.5	87.5	87.5	25	12.5	0
	Short-I.S.	37.5	37.5	50	50	50	50	75	62.5	50	50
Wh-whether	Short-N.I.	87.5	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Long-N.I.	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	87.5	87.5
	Short-I.S.	62.5	**100**	**100**	**100**	**100**	**100**	25	62.5	**100**	**100**
Average		70.8%	79.2%	**88.5%**	84.4%	84.4%	83.3%	66.7%	62.5%	62.5%	61.5%

Table 4.2: Accuracy results for Gpt-2 and BERT Italian models, on the Italian test suite from Sprouse et al. (2016).

Model	**Adjunct**	**Complex NP**	**Wh**
Gpt2 (Warstadt et al 2020)	91	72	77
Gpt2-large	90.2	72	79.1
Gpt2-medium	**91.6**	72.3	77.9
Gpt2	91.3	68.8	82.2
BERT-large-cased (PenLP)	86.3	67.4	69.4
BERT-base-cased (PenLP)	88.1	56	66.2
RoBERTa-large (PenLP)	86.9	**82.3**	**88.2**
RoBERTa-base(PenLP)	84.9	75.3	76.2
RoBERTa-base(PLL)	80	47.7	83.3

Table 4.3: Accuracy results on some extraction islands phenomena in the BLiMP English test suite. The score on the first row are taken from the original paper Warstadt et al. (2020)
NB: RoBERTa-base with alpha = 0.8 in the penalty term as in Lau et al with with alpha = 1 ..

4.2 Factorial tests results

4.2.1 Discussion on plots from Gpt-2, softmax and PenLP

In this section we discuss the results and plots of the scores obtained from the Gpt-2 model outputs with the softmax activation function and the PenLP sentence acceptability measure.

We choose to focus on this combination because it seems to produce in general more accurate results as seen in Table 4.1.

We include in the appendix the plots for the BERT models and the other acceptability measures (LP and PenLP based either on the model outputs after softmax of logistic activation functions).

..

Complex np islands

In the middle and right images on Figure 4.1, we see that, for complex NP items, the long non-island sentences on average get scored by the Gpt-2 model with a lower acceptability than by the human subjects (left image). This seems to be due to the long distance dependency effect, that gets exacerbated as a sentence increases in lenght (in these test suites, the complex NP islands examples have longer sentences).

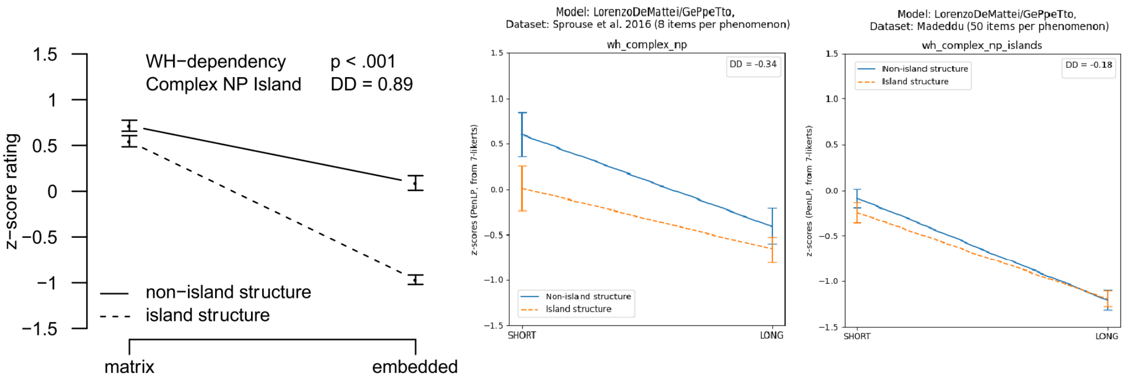


Figure 4.1: Comparison of plots for wh-dependencies complex NP islands

The first plot on the left shows the scores on humans subjects published in Sprouse et al. (2016) for Italian complex NP islands with wh-dependencies. For each line, the left-most edge represents the score for the short-distance dependency sentence, the right-most the long-distance dependency. The plot in the middle shows the scores from a Gpt-2 model (De Mattei et al., 2020) on the same test suite used for the first plot. The plot on the right shows the scores from the same Gpt-2 model but on the expanded test suite developed for the present thesis.

Doing some variations experiments (see Table 4.4), we found that by replacing the main clause verbs with "intuire che"/"avere l'intuizione che" ("to sense that" / "to have the intuition that") or "avvertire che"/"avere il sentore che" ("to feel that" / "to have an inkling that") or "percepire che"/"avere la percezione che" ("feel that"/"to have the feeling that"), the island restriction violation seem to have no effect, and the Long-Island sentence becomes more acceptable than the same sentence without the island structure (Long-NonIsland).

Indeed, the sentence *Cosa hai avuto l'intuizione che il portavoce avrebbe confermato?* (*'What did you have the intuition that the spokeperson had confirmed?'*), seems acceptable despite extracting from a complex NP construct.

An hypothesis, whose demonstration we leave for future work, is that this is due to main clause expressions in which the subject has the semantic role of an EXPERIENCER rather than an AGENT, which could be a condition for Complex NP island restrictions to enter into effect or not.

With other variations experiments, we found that replacing the main clause verb with "sapeva che"/"conosceva che" ("he knew that", see examples in table ..TODO), increases the acceptability of the long-non island sentences; on the other hand, this variation results in a lower acceptability to the short island sentence, compared to the scores from human subjects, and by this way the non-island and island line end up being almost parallel, with the DD score close to zero (which indicates an almost absent island effect).

Long-NonIsland		Long-Island		
text	**PenLP**	**text**	**PenLP**	**Diff**
Cosa hai messo in dubbio che il portavoce avrebbe confermato?	-31.65	*Cosa hai messo in dubbio la previsione che il portavoce avrebbe confermato*	-33.21	1.56
Cosa hai intuito che il portavoce avrebbe confermato?	-32.36	*Cosa hai avuto l'intuizione che il portavoce avrebbe confermato?*	-29.99	-2.37
Cosa hai detto che Gianni avrebbe sollevato?	-33.74	*Cosa hai riferito il fatto che Gianni avrebbe sollevato?*	-36.64	2.90
Cosa hai intuito che Gianni avrebbe sollevato?	-34.31	*Cosa hai avuto l'intuizione che Gianni avrebbe sollevato?*	-30.52	-3.79
Cosa hai messo in dubbio che io avrei vinto?	-26.48	*Cosa hai messo in dubbio la previsione che io avrei vinto?*	-30.17	3.69
Cosa hai intuito che che io avrei vinto?	-29.49	*Cosa hai avuto l'intuizione che io avrei vinto?*	-24.45	-5.04

Table 4.4: Comparing acceptability variations among sentences in the complex NP dataset. A positive difference indicates that the Long-NonIsland sentence is more acceptable than the Long-NonIsland one, as expected.

Whether islands

From Figure 4.2 for whether islands with wh-dependencies, we can see that the Gtp-2 model, with the PenLP sentence acceptability estimate, compared with the results from humans gives higher scores for all sentence types. This difference is more pronounced in the scores performed by Gpt on the original test suite from Sprouse et al. (2016) (middle image).

The slope of the lines (both for island and non-island sentence structures) is however quite similar to the human scores.

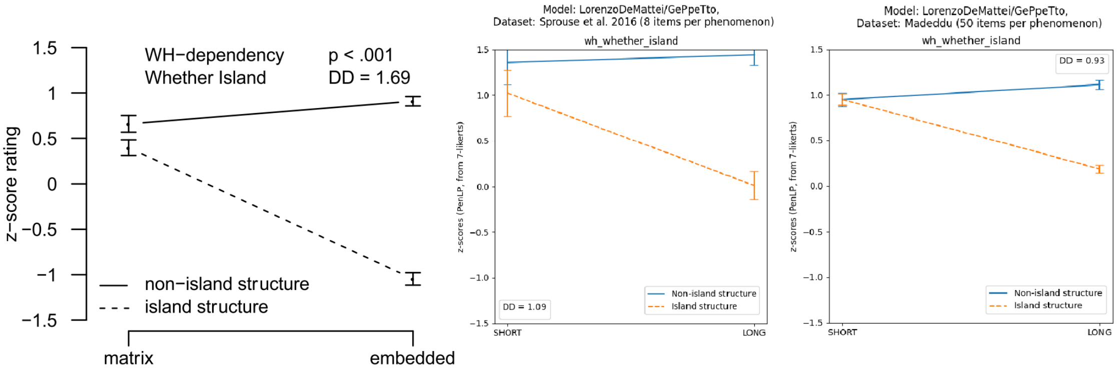


Figure 4.2: Comparison of wh-dependency whether islands

..

Experimenting with other sentence variations for whether islands, we found that a variation that aligns the plots more to those from the humans' scores, is replacing the personal pronouns (like "io") with proper nouns (like "Gianni") or common nouns (like "il parlamentare", or "lo studente") like in this example:

Example 4.2.1. Whether islands, Long-Island sentence type

a. Cosa ti domandi se io abbia riscosso?

b. Cosa ti domandi se Gianni abbia riscosso?

c. Cosa ti domandi se il parlamentare abbia riscosso?

TODO: test sentence variations using less common verb tenses and moods (but consistently across an item or even all items of a suite)

Subject islands

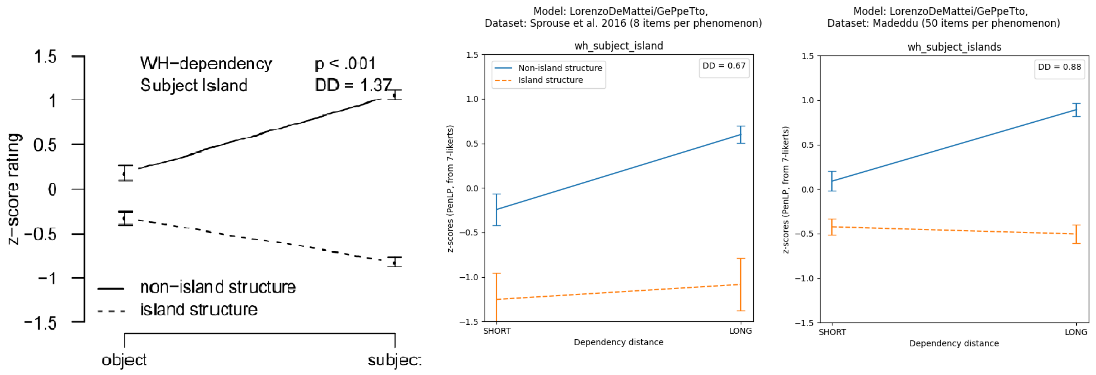


Figure 4.3: Comparison of wh-dependencies subject islands

In Figure 4.3, the middle image of Gpt scores on Sprouse data shows significantly lower scores (compared to the human scores on the left image) for the sentences with an island structure (dotted line). However, on the right image (Gpt scores on our test suite) they are similar to the human scores.

For the non island line, the Short-NonIsland is scored lower then on human subject by Gpt for both testsuites. The Long-NonIsland scores instead are similar (around 1.0). The slope of the lines on the right image is similar to the one for the human scores (left image).

Adjunct islands

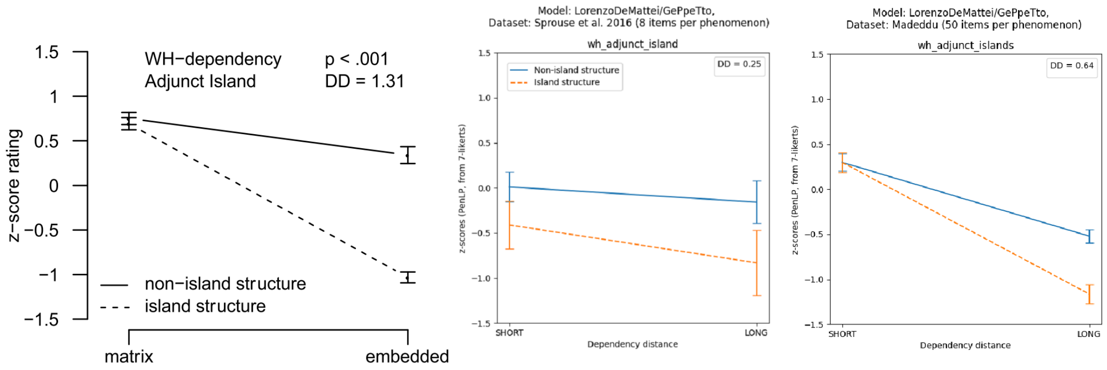


Figure 4.4: Comparison of wh-dependencies adjunct islands

In Figure 4.4 the slope of both lines on the new test suite (right image) is similar to the human scores (left image), althoughg with a lower average acceptability for the Long-NonIsland sentences, which results in a downward steeper line.

Note that in the test suite we developed for this thesis, the Long-Island sentences for adjunct islands have a form that might make them more easily identifiable has unacceptable than the ones in the Sprouse test suite: compare *Che cosa Gianni è partito per Parigi dopo aver fatto?* (new dataset) with *Cosa ti irriti se dimentico in ufficio?*. The ending of the first sentence might be less frequent and result in a lower acceptability score.

Overall observations across all four island phenomena

The Gpt-2 scores on the original Sprouse et al. test suite (middle image) have wider standard error bars, which are considerably smaller for the new test suite developed for the present thesis (right image).[1]

While on human ratings the unacceptable sentences (Long-NonIslands) receive all on average a z-score of about -1, there is more variability in the scores given by the Gpt model, in particular for whether islands sentences (Figure 4.2), which receive much higher acceptability rating on average (betwenn 0 and 0.5 the Sprouse test suite and ours). ..

4.2.2 Discussion on plots from BERT, the logistic function, and PenLP

In Figure 4.5 we see the plots of the scores from the BERT model, with the PenLP-L sentence acceptability estimate (using the logistic function rather

[1] This might be due to the fact that the number of items between the two test suites increases from 8 to 50.

than the softmax, as described in ??). Interestingly, it seems that this combination of BERT with the logistic function makes a clear separation between unacceptable sentences (the Long-NonIsland type) and acceptable sentence types, with minimal, if any, difference in score between acceptable sentences.

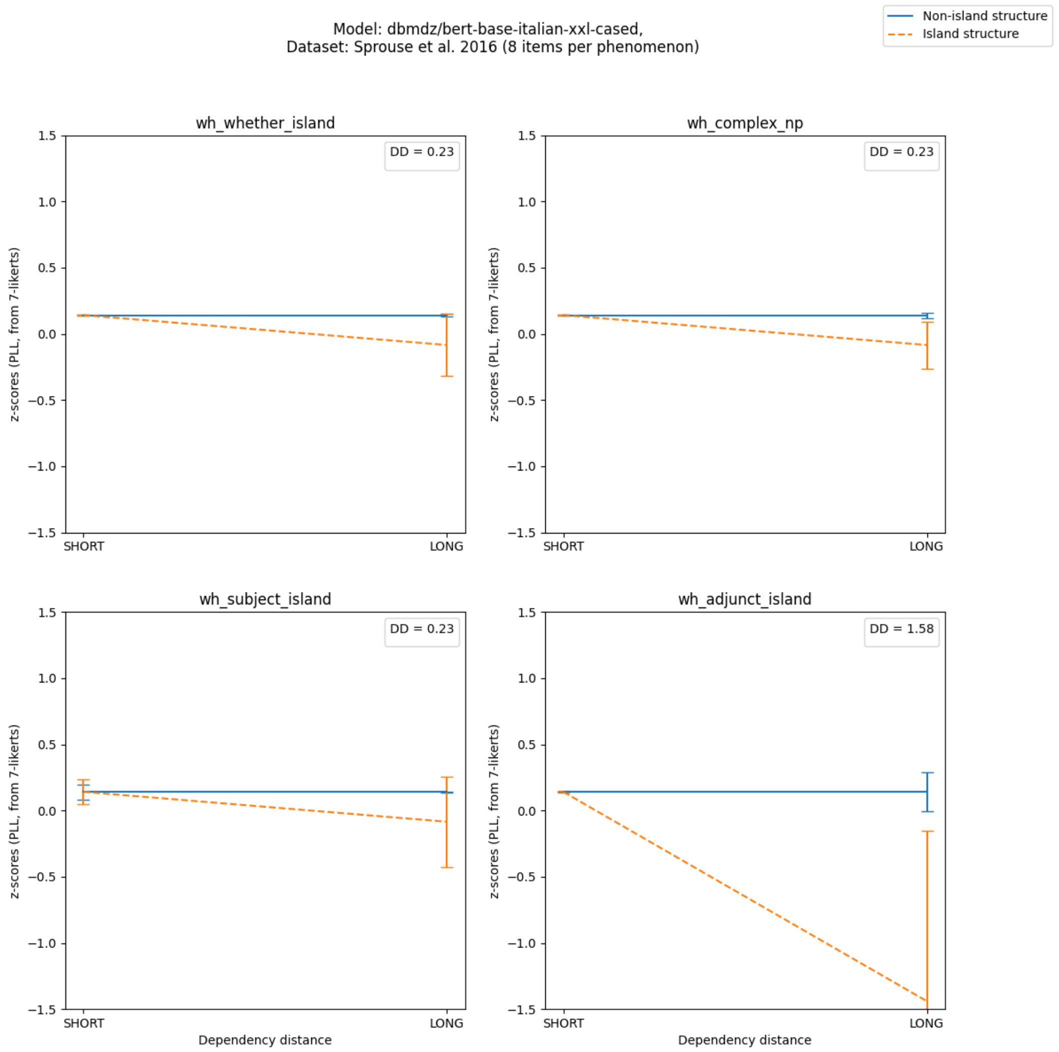


Figure 4.5: Plots from BERT with the PenLP-L sentence acceptability approximation

If we compare these plots with the accuracy scores in Table 4.1 from this model (BERT, logistic function, PenLP) , we see that they are among the best with an average accuracy of 80.7%; however, looking at the scores for each phenomenon, the scores are high only in about half the cases, and quite low for the rest: Long-NonIsland sentences of all four island phenomena are scored with 58%, 64%, 74%, 58% accuracy, and the Short-NonIsland of subject islands with 28% accuracy.

4.2.3 Discussion on plots from GilBERTo

Overall, for GilBERTo, the plots are significantly different from those on human ratings. (..)

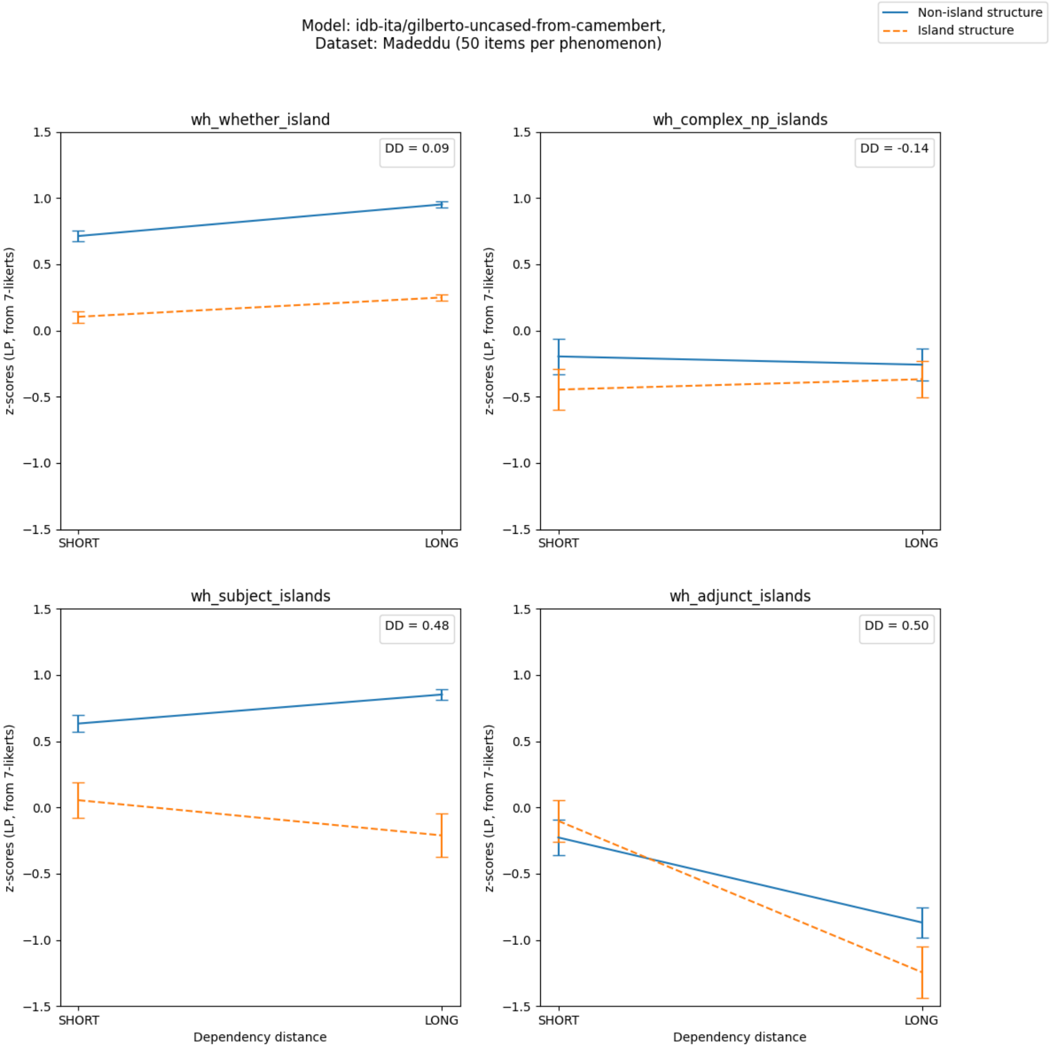


Figure 4.6: Plots from GilBERTo with the LP sentence acceptability approximation

4.3 Draft notes on the results

4.3.1 Observation on plots for other models

The plots for Geppetto and PenLp are the most similar to the plots on human ratings. However, with the LP measure, the plots differ significantly.

The plots for BERT with PenLP on the Sprouse data are similar to the results on humans from Sprouse et al. (2016),

In the plots from the GilBERTo scores, we see that for complex NP and subject islands, the sentences with the island structure receive a higher acceptability than the non island ones (this is no longer the case when removing the penalty for sentence lenght, using the LP sentence acceptability measure). With the PenLP sentence acceptability approximation, the lines tend to have similar slopes, close to being parallel (indicating a lack of, or small, island effect).

We refer to the Appendix for the plots for the other models.

4.3.2 What seems to affect the models acceptability scores

Adjunct islands

Guardando le short-nonislands, sembra anche qui preponderante il fatto di usare nomi propri di persona (che ha uno score di accettabilità minore) e usare invece di nomi comuni animati/di mestieri/.. (che aumenta l'accettabilità). Ma questo non sembra influenzare il DD score finale

4.3.3 Discarded observations on differences between the plots

Other notes on Complex np islands, what seems to affect the models acceptability scores

In Figure 4.1 we see that the non-island line (the line connecting the two acceptability scores for short and long distance dependency sentences without an island structure) is significantly lower in new data (todo: see constructs that increase the score of both long and short)

The (gpt) model seem not to make much difference in ..acceptability btw "regular" subordinates and complex noun phrases ..

– also the short island point is significantly lower

All the 50 + 8 items of the two test suites (the ones from Sprouse et al and the ones developed for the present thesis), when altered to use the following construct for the complex NP "avuto l'intuizione che" (had the intuition that), are scored by the model as if there is no island restriction violation anymore.

il verbo "intuire" diminuisce ..l'accuratezza del giudizio di accettabilità in particolare diminuzione della accettabilità delle frasi long nonisland (con long distance wh dependency e struttura non island), che ricevono accettabilità minore di quelle con struttura island: Esempio analisi variazioni con verbo "intuire" complex np, ..

the verb form "sapeva" (imperfect), compared to ..present perfect forms ("ha osservato/affermato/..") seems to get better DD score in complex np islands (and also in another phenomenon ..).

Whether islands

NB: note that according to human scores, for this type of whether island sentences, the "correct" scoring is to have an increase in acceptability going from short to long non island sentences. Maybe comparing the scores between acceptable sentences (in this case short and long non islands), expecting it to match humans acceptability judgements .. is beside the present ..research question. In any case, we noted a reverse in the acceptability difference between this two types of sentences (short and long non islands) when changing the subject of the subordinate sentence from a personal pronoun ("io", 1st pers sg), to a proper noun (i.e. "Gianni"), to a common noun (i.e. "il parlamentare").

4.4 BLiMP English dataset

..

4.4.1 English models details

..

4.5 Token surprisal analysis

(TODO)

4.6 Follow up experiments and future work

(..)

Chapter 5

Conclusions

..

Appendix A

Factorial design plots

Scores obtained from the following Huggingface pretrained Italian models:

BERT: dbmdz/bert-base-italian-xxl-cased [1]
GilBERTo: idb-ita/gilberto-uncased-from-camembert [2]
GePpeTto: LorenzoDeMattei/GePpeTto (De Mattei et al., 2020) [3]

[1] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

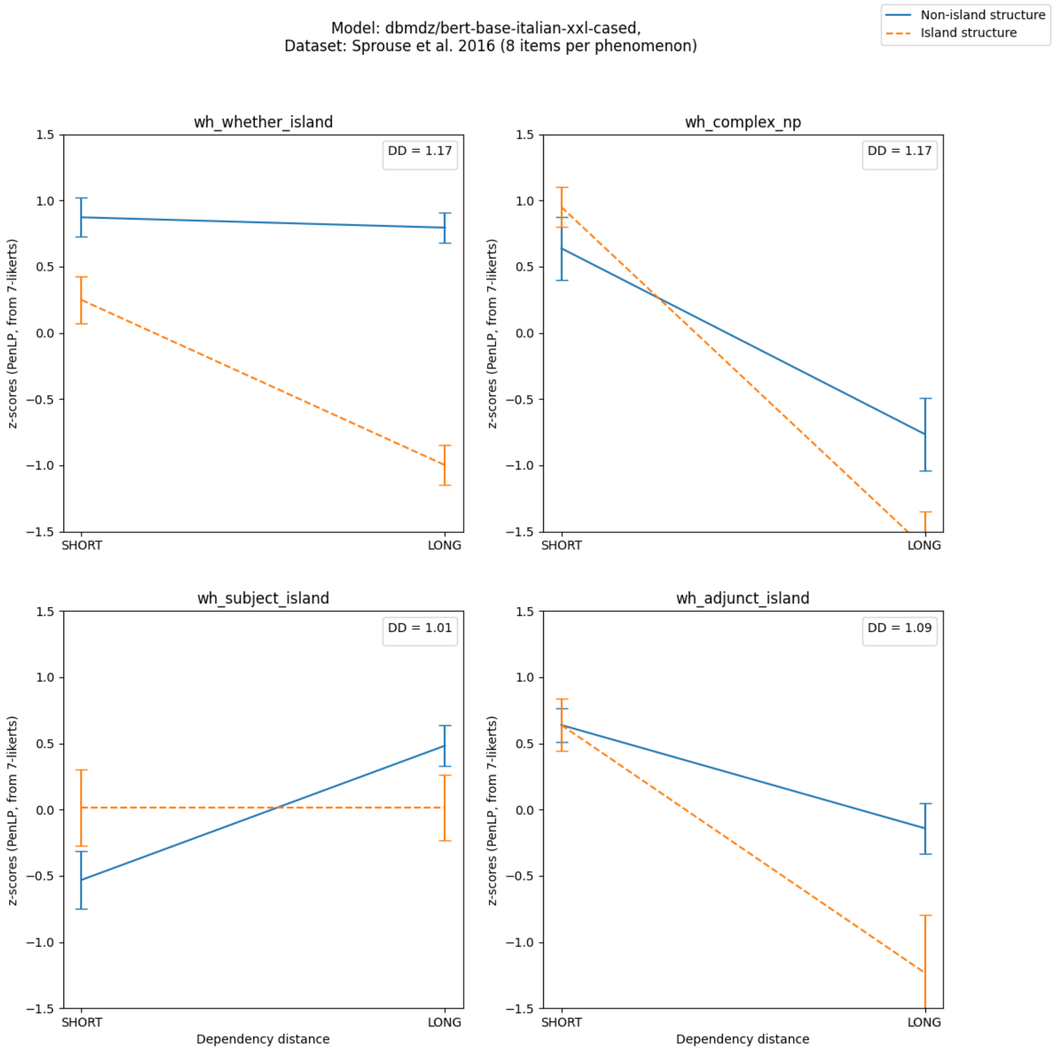
[2] https://huggingface.co/idb-ita/gilberto-uncased-from-camembert

[3] https://huggingface.co/LorenzoDeMattei/GePpeTto

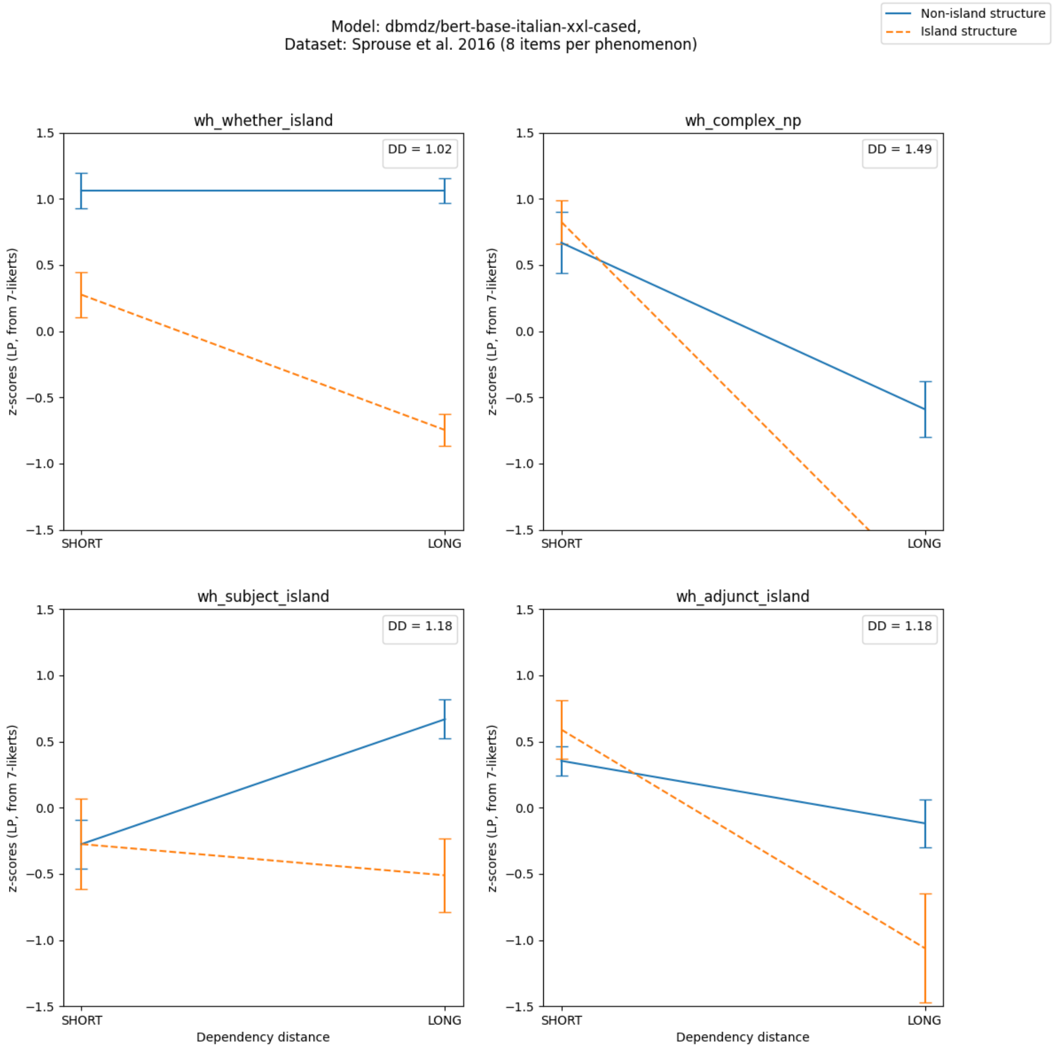
A.1 Sprouse test suite

A.1.1 BERT

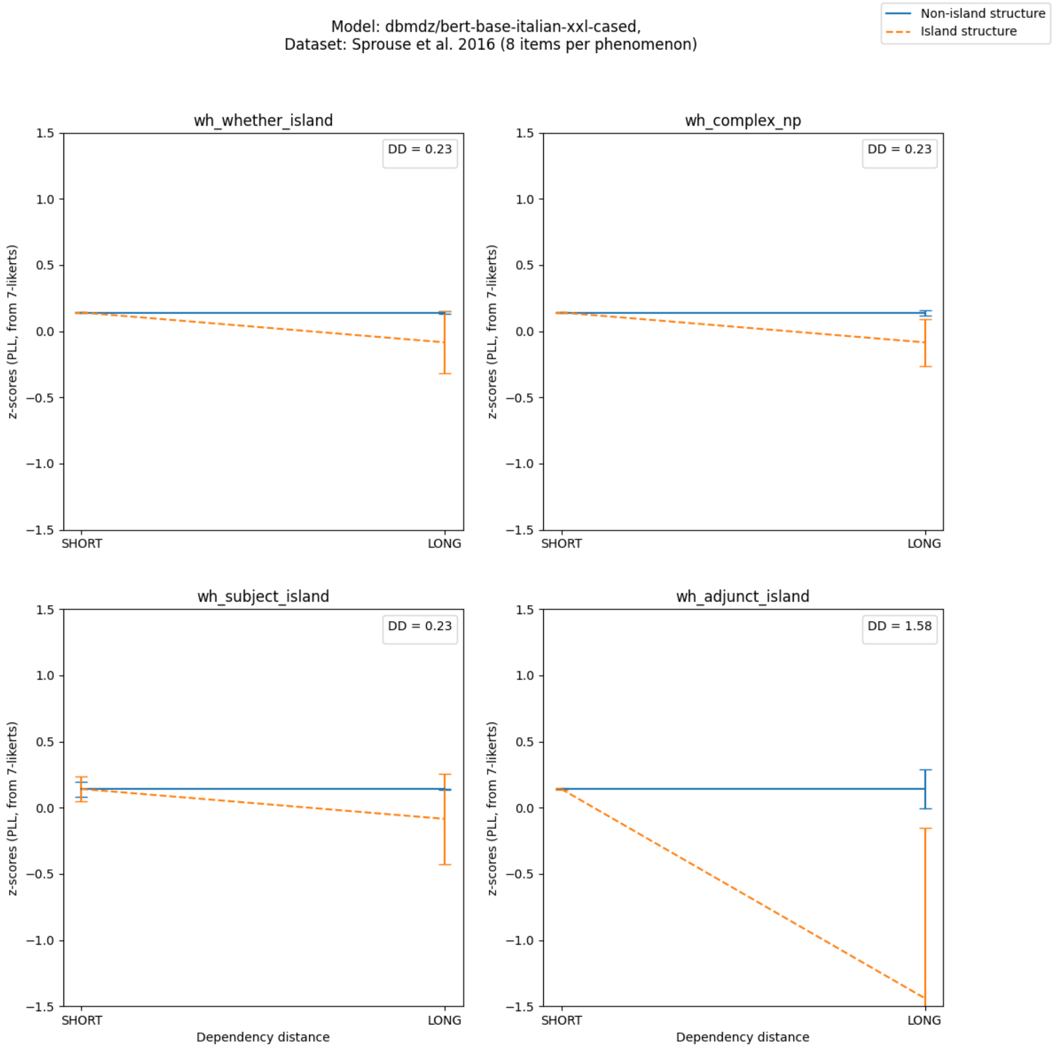
BERT with PenLP (from softmax model output)



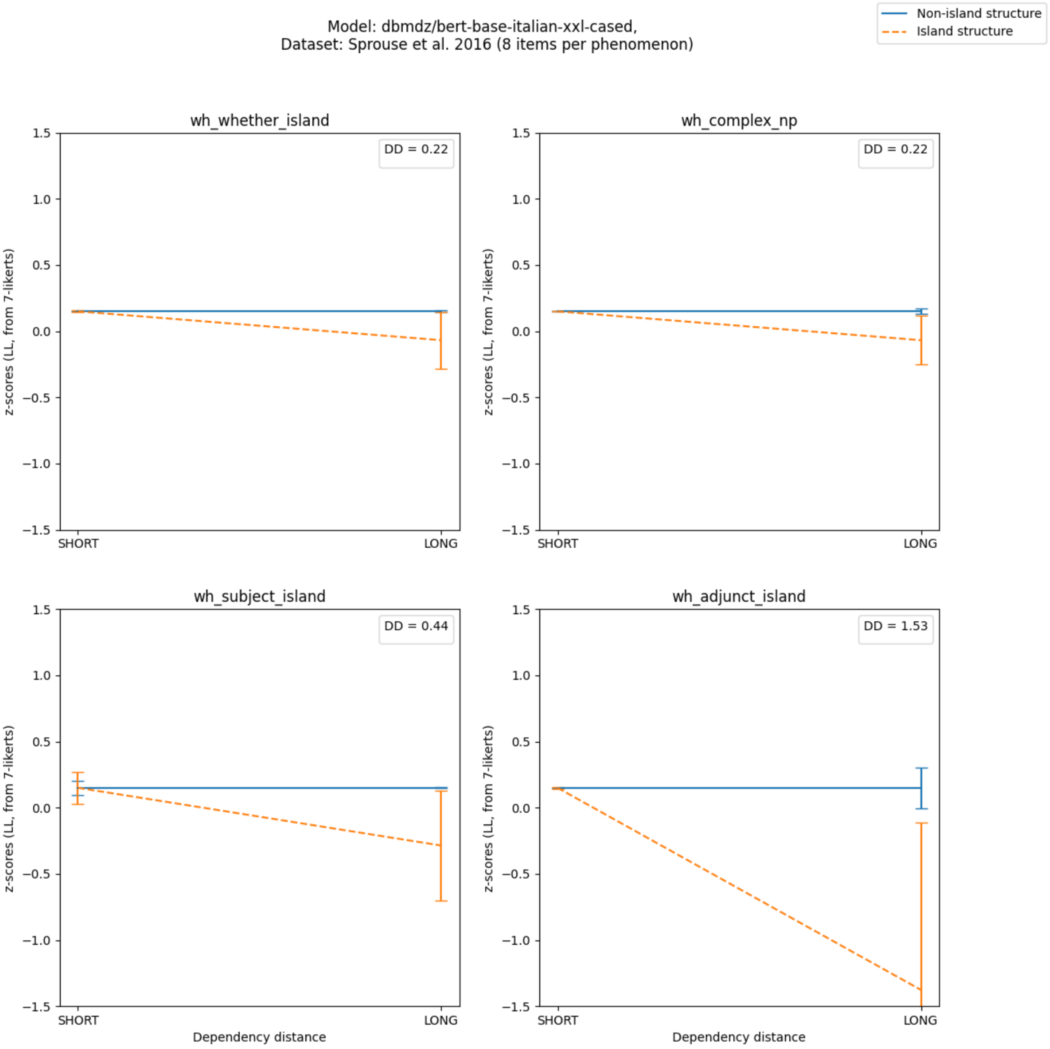
BERT with LP (from softmax model output)



BERT with PenLP-L (from logistic function model output)

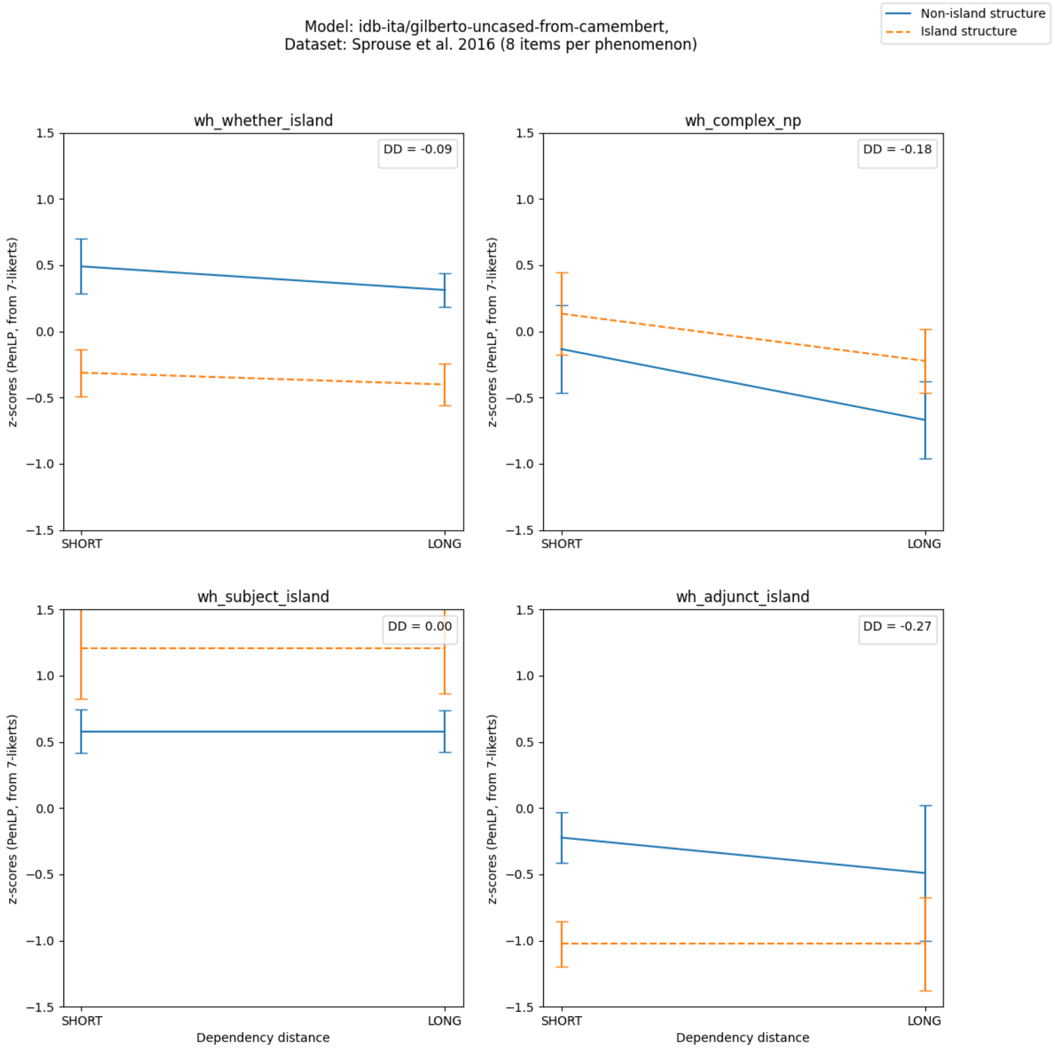


BERT with LP-L (from logistic function model output)

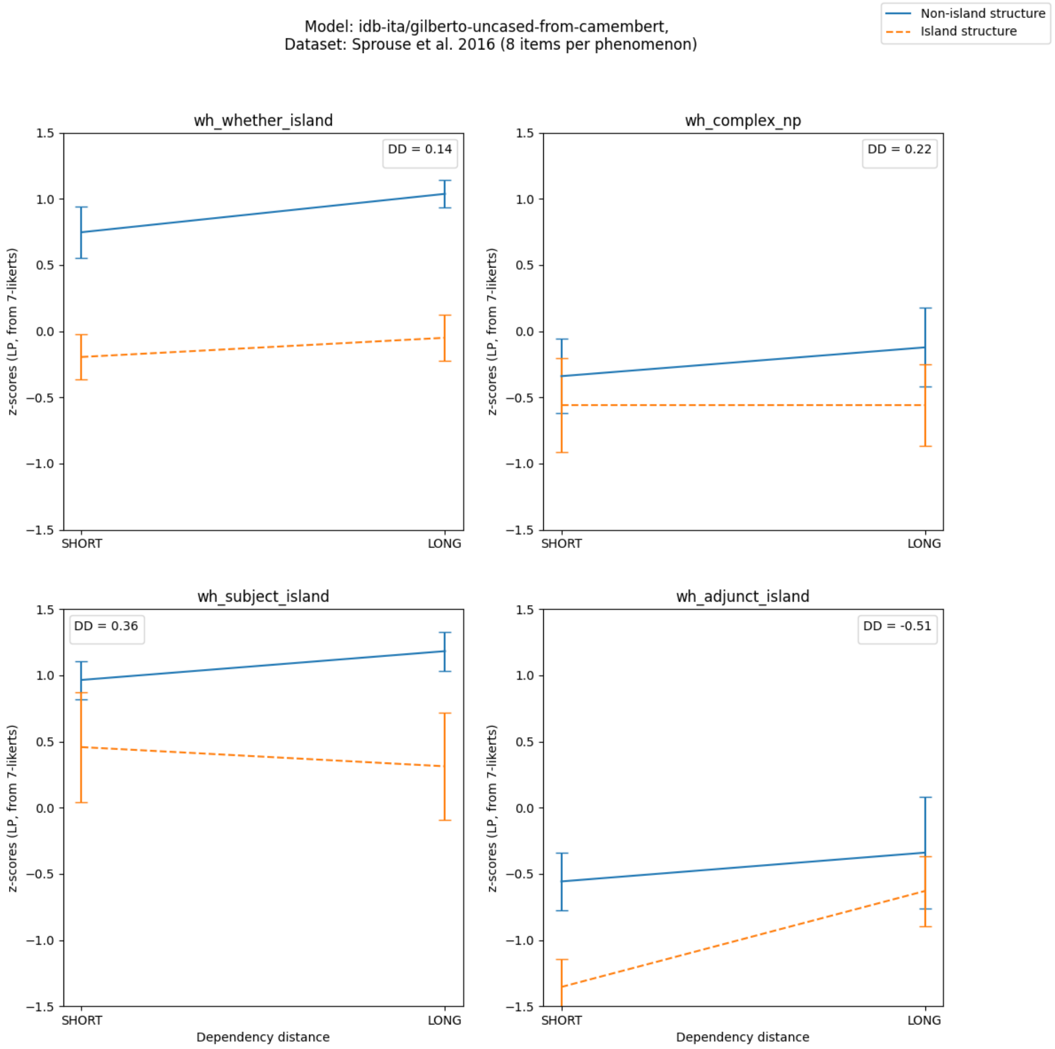


A.1.2 GilBERTo

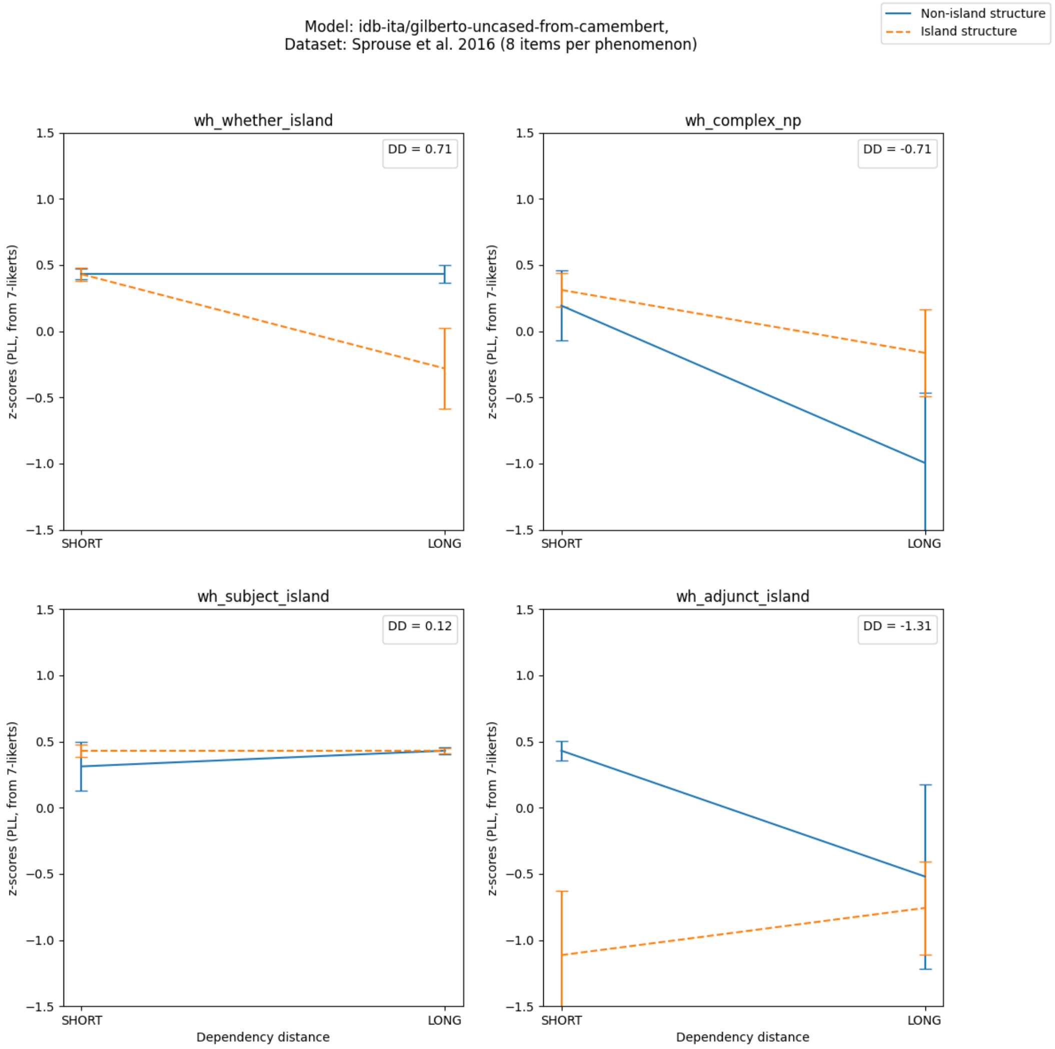
GilBERTo with PenLP (from softmax model output)



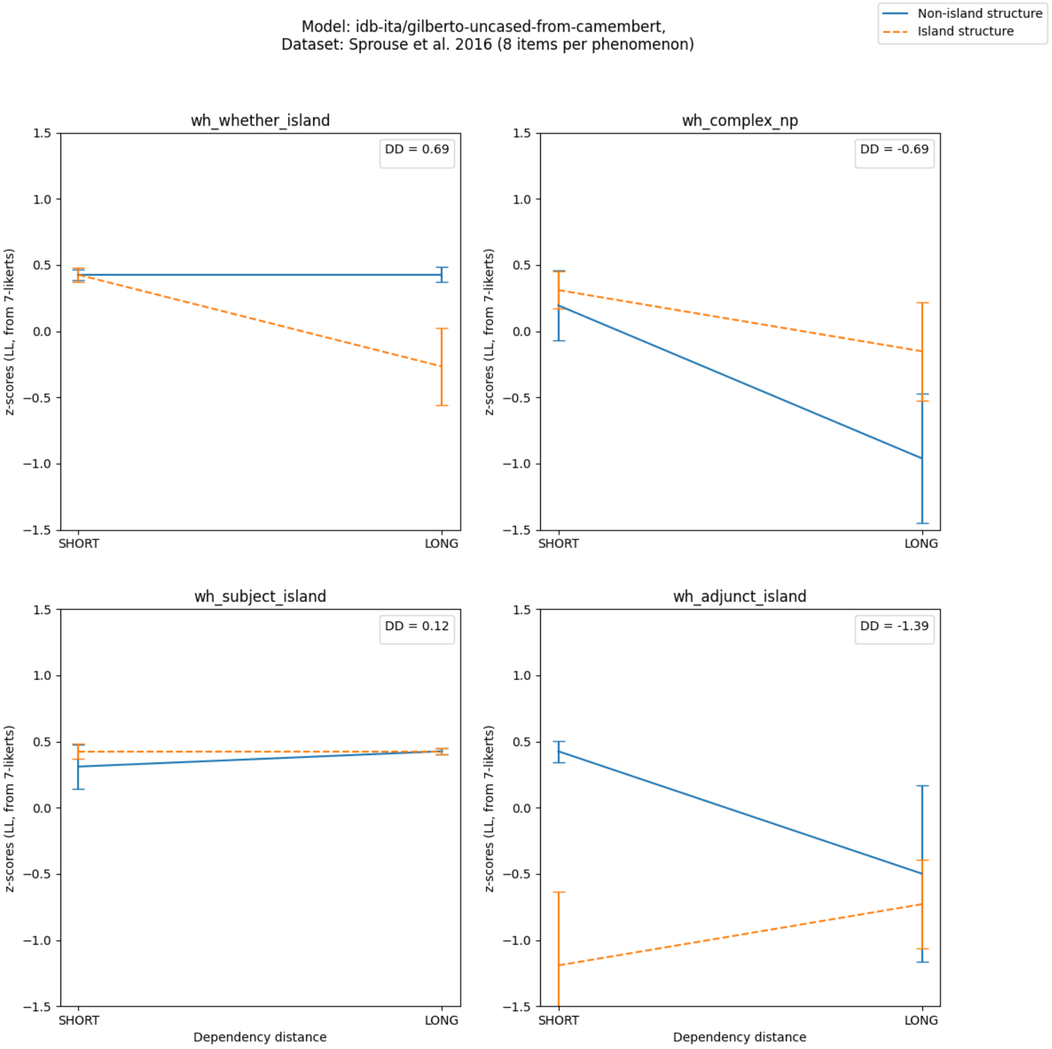
GilBERTo with LP (from softmax model output)



GilBERTo with PenLP-L (from logistic function model output)

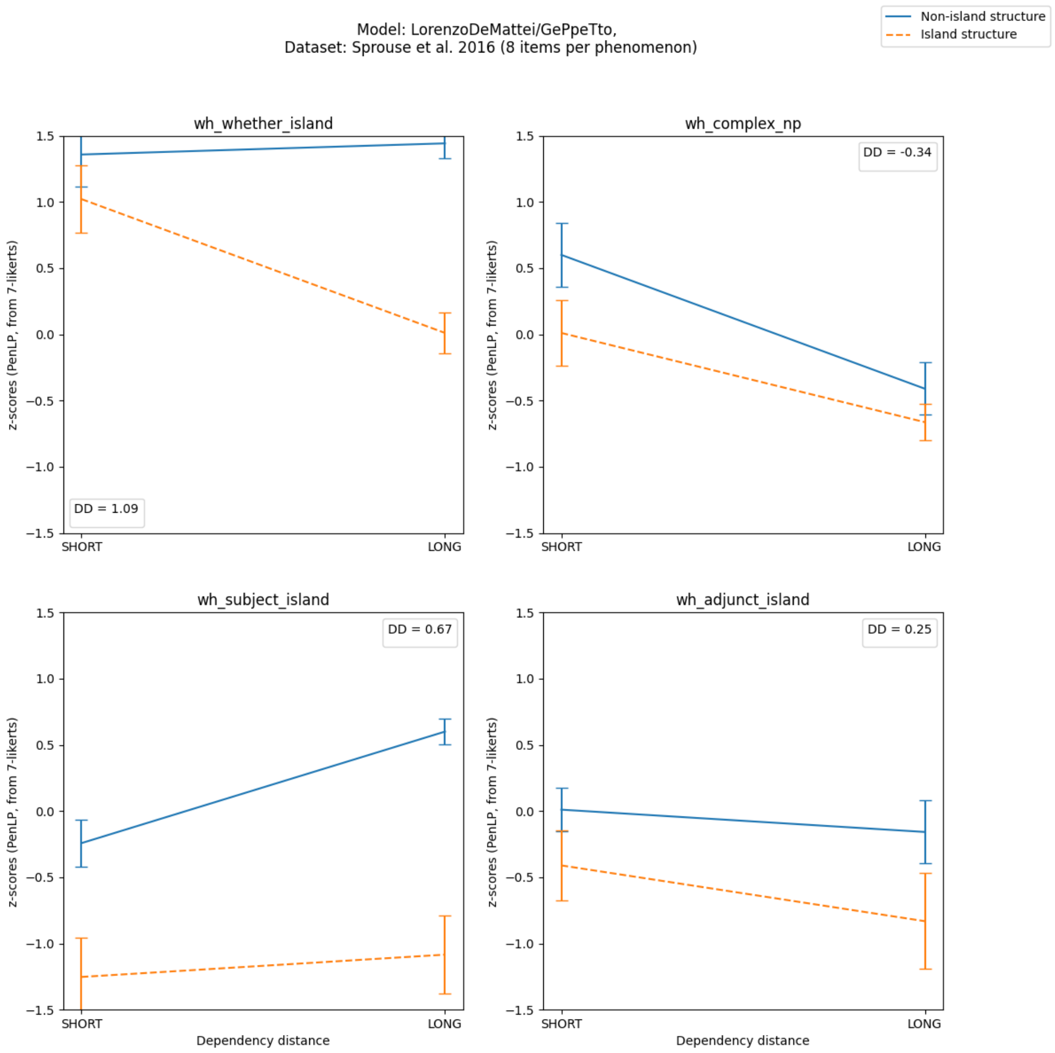


GilBERTo with LP-L (from logistic function model output)

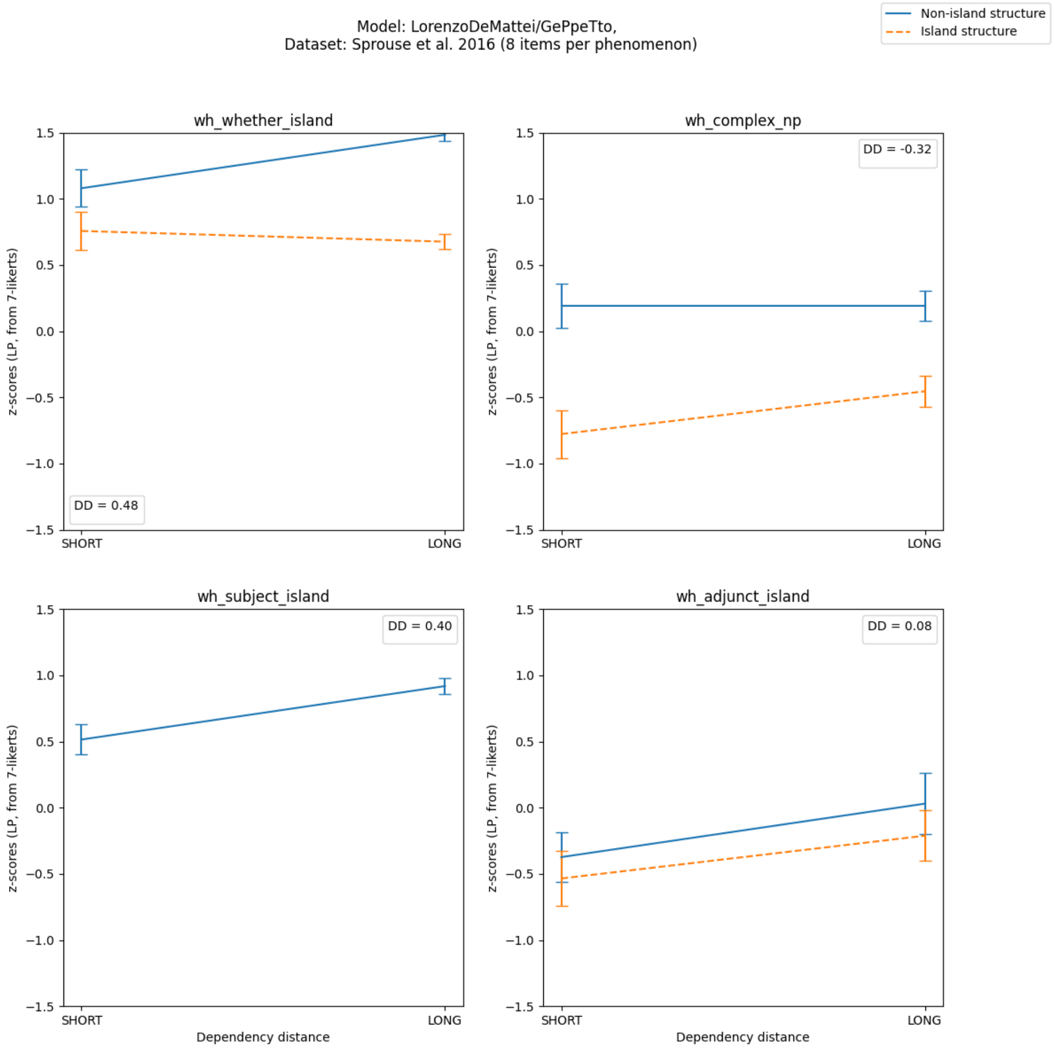


A.1.3 GePpeTto

GePpeTto with PenLP (from softmax model output)



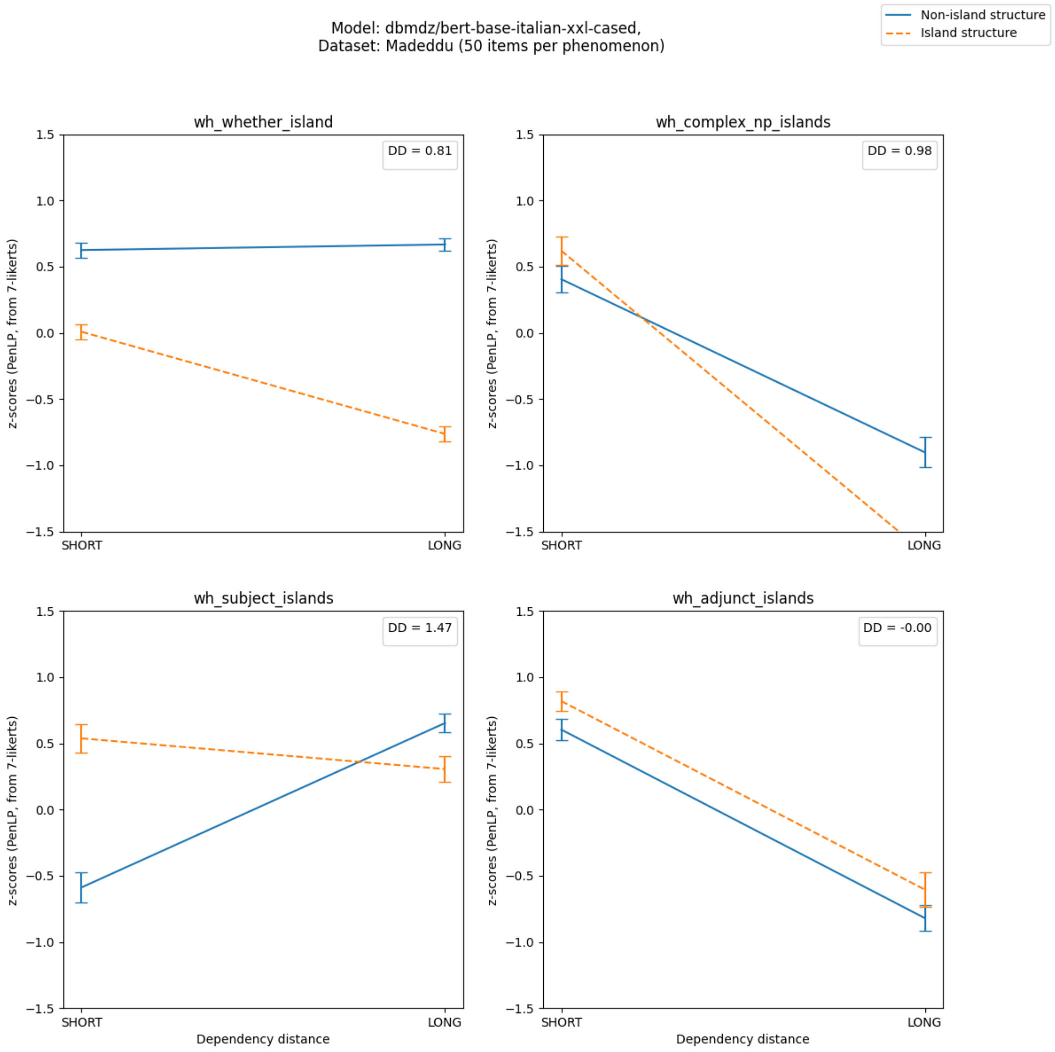
GePpeTto with LP (from softmax model output)



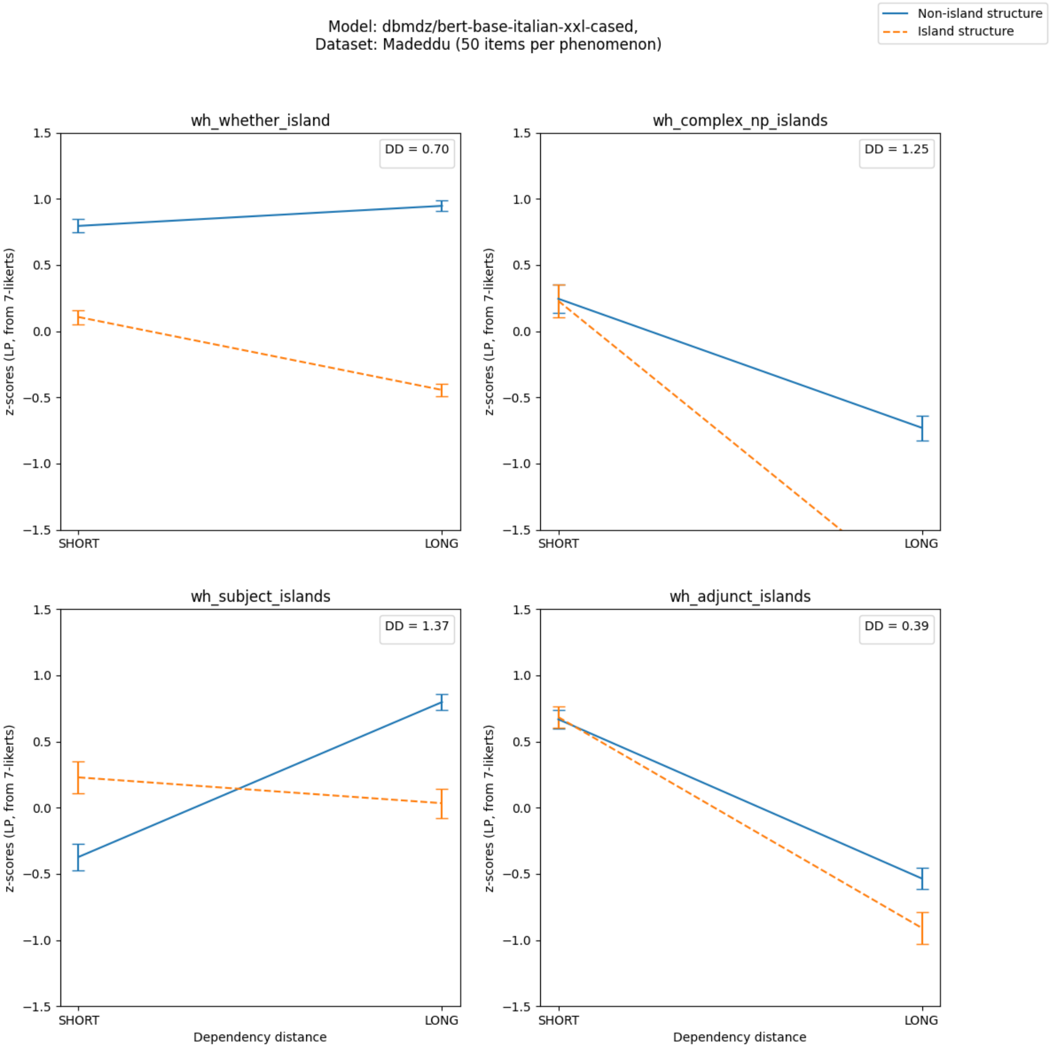
A.2 Madeddu test suite

A.2.1 BERT

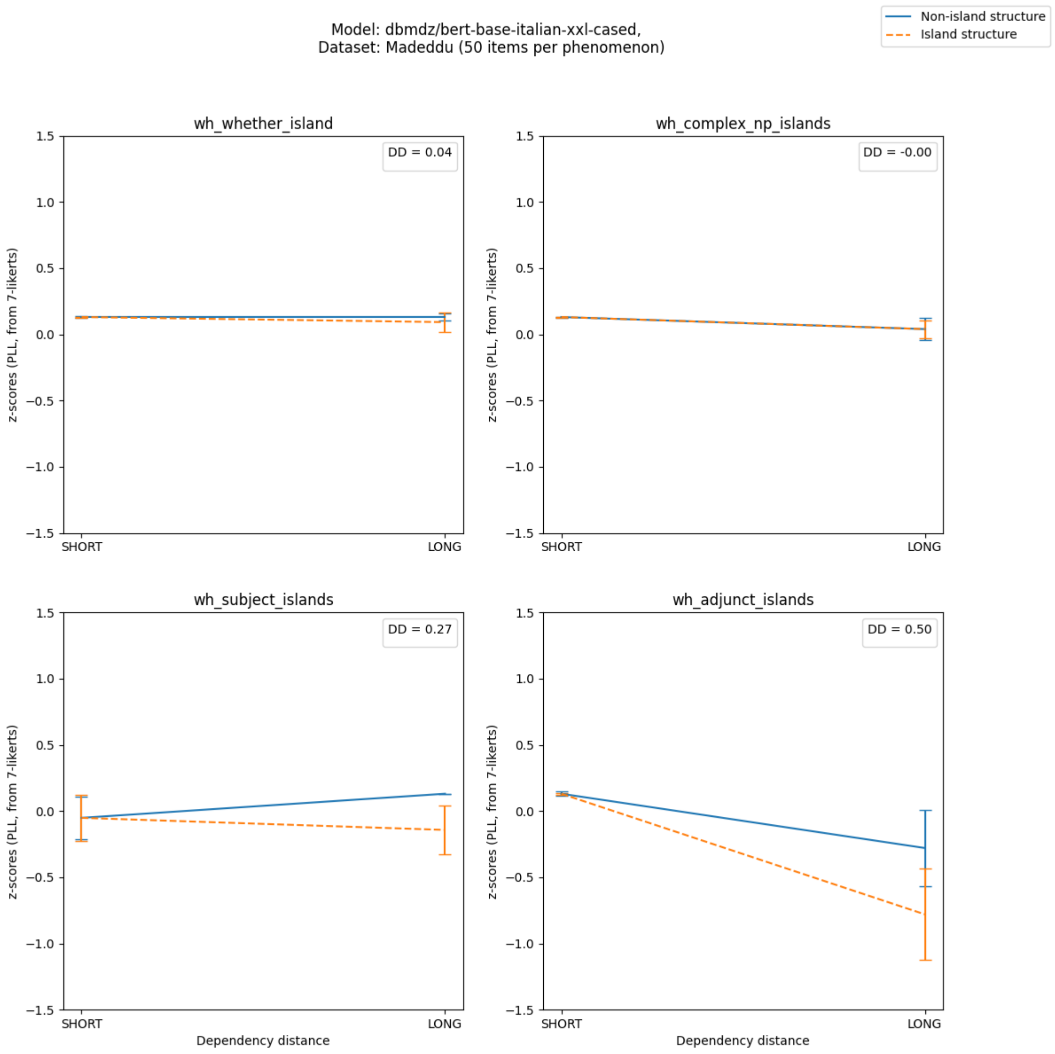
BERT with PenLP (from softmax model output)



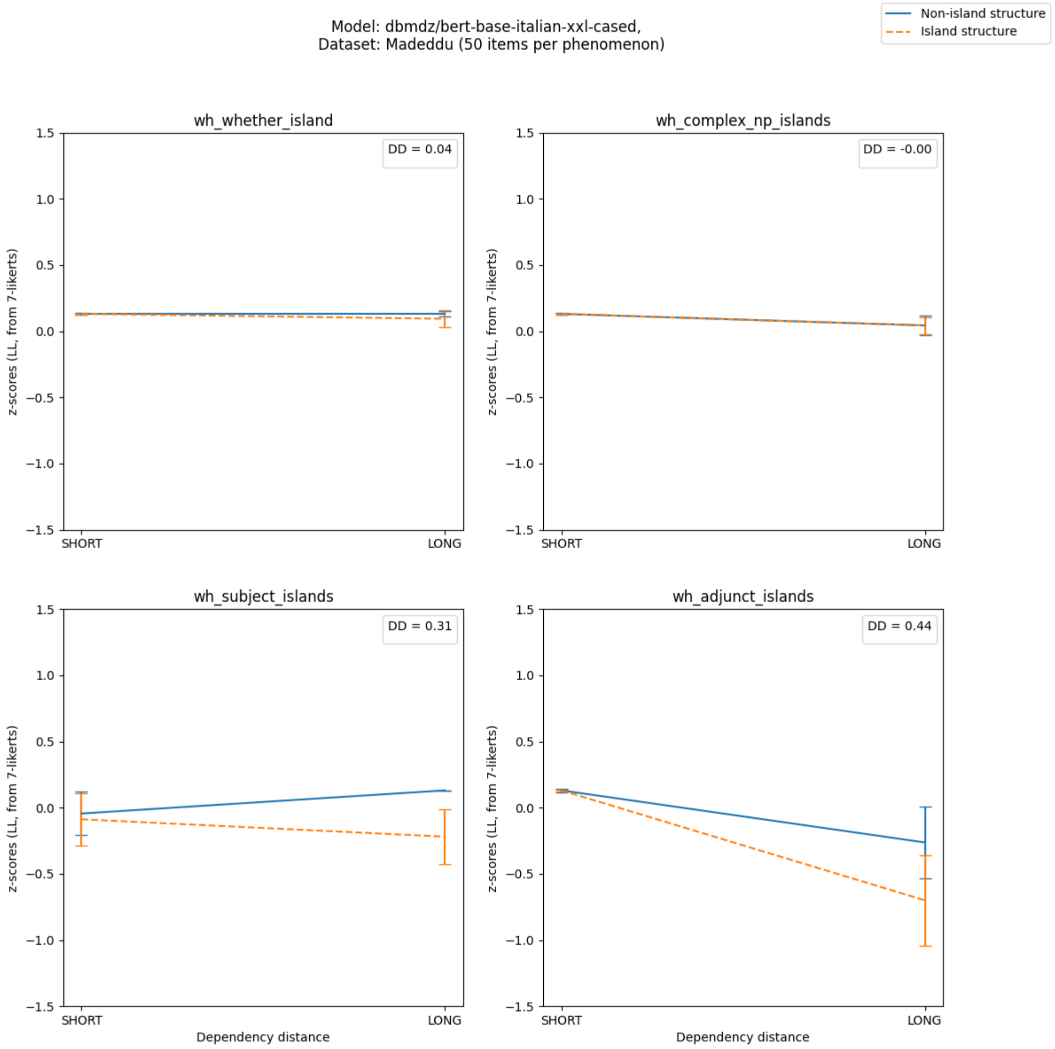
BERT with LP (from softmax model output)



BERT with PenLP-L (from logistic function model output)

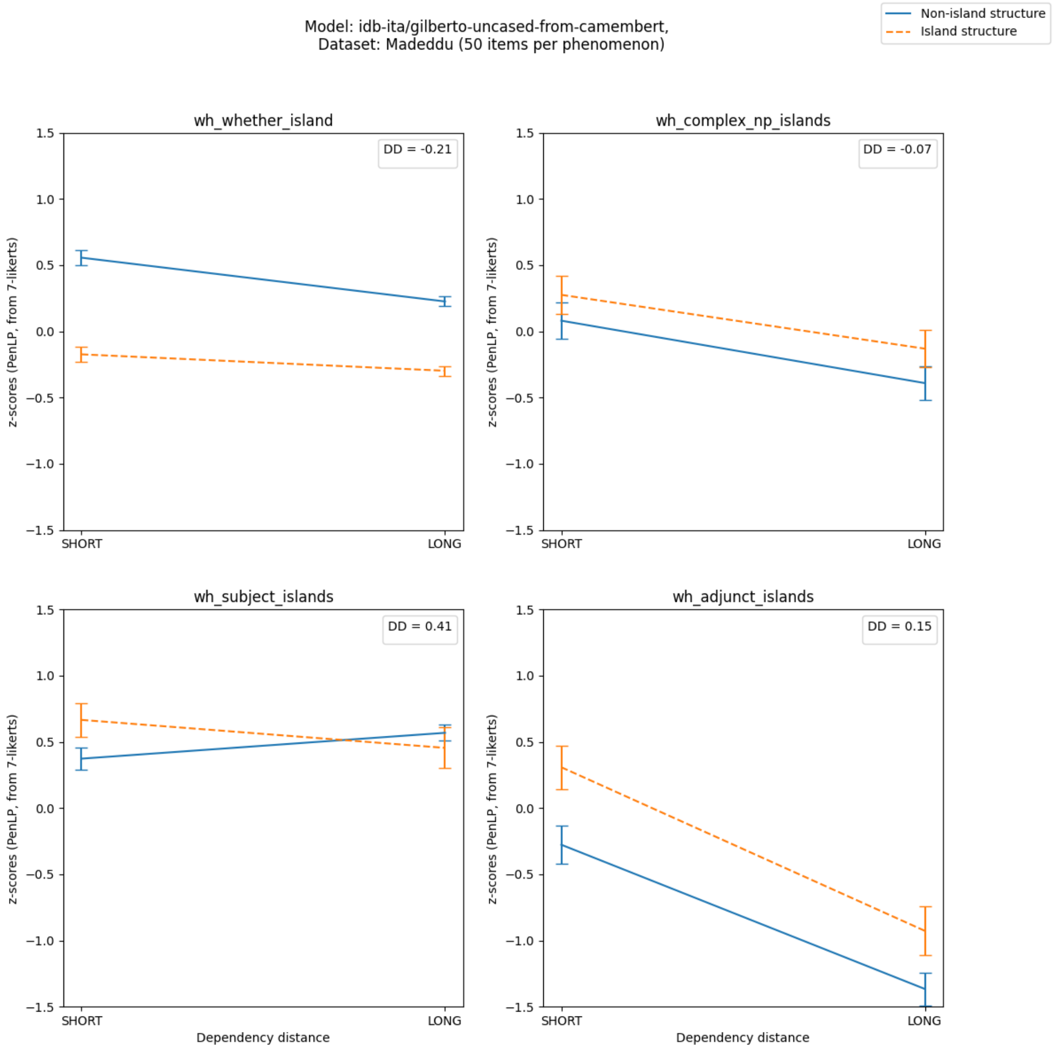


BERT with LP-L (from logistic function model output)

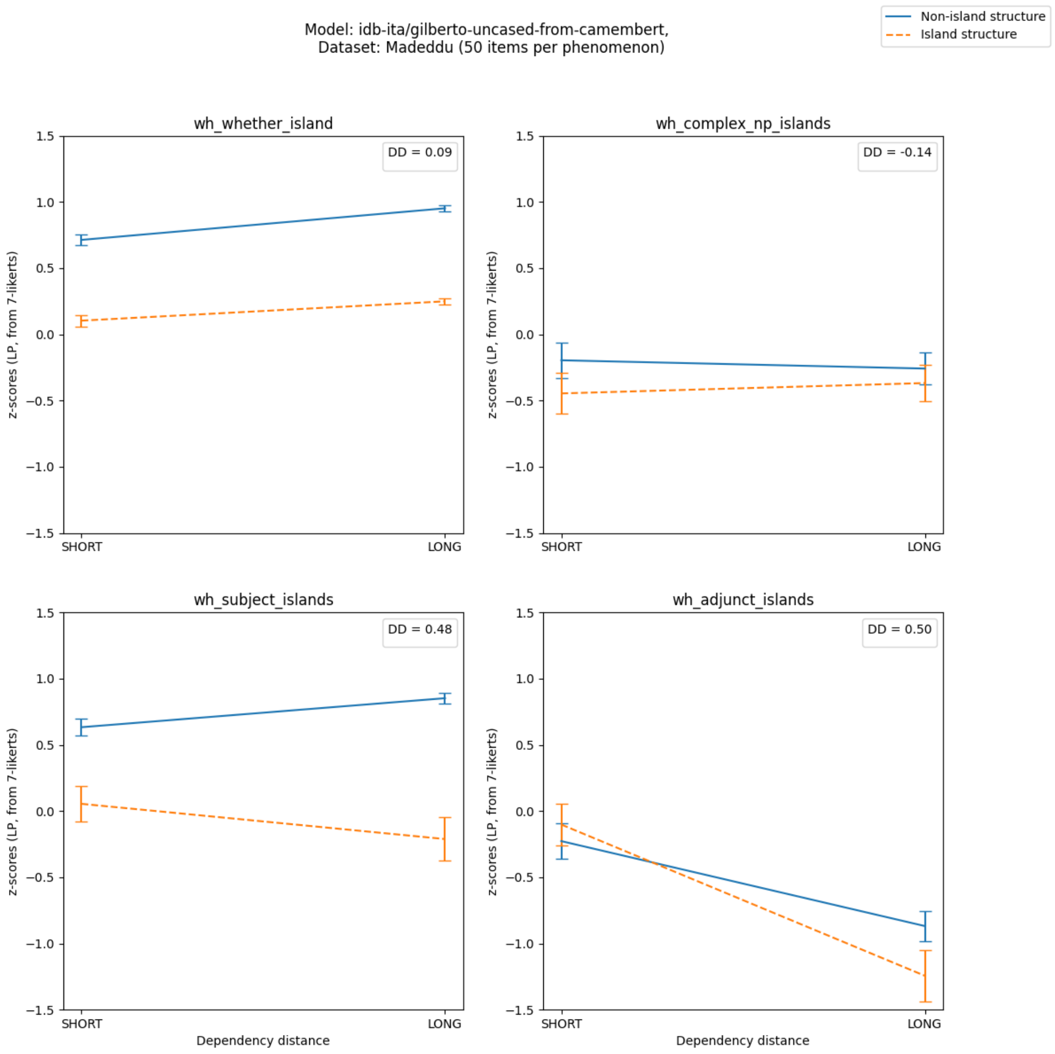


A.2.2 GilBERTo

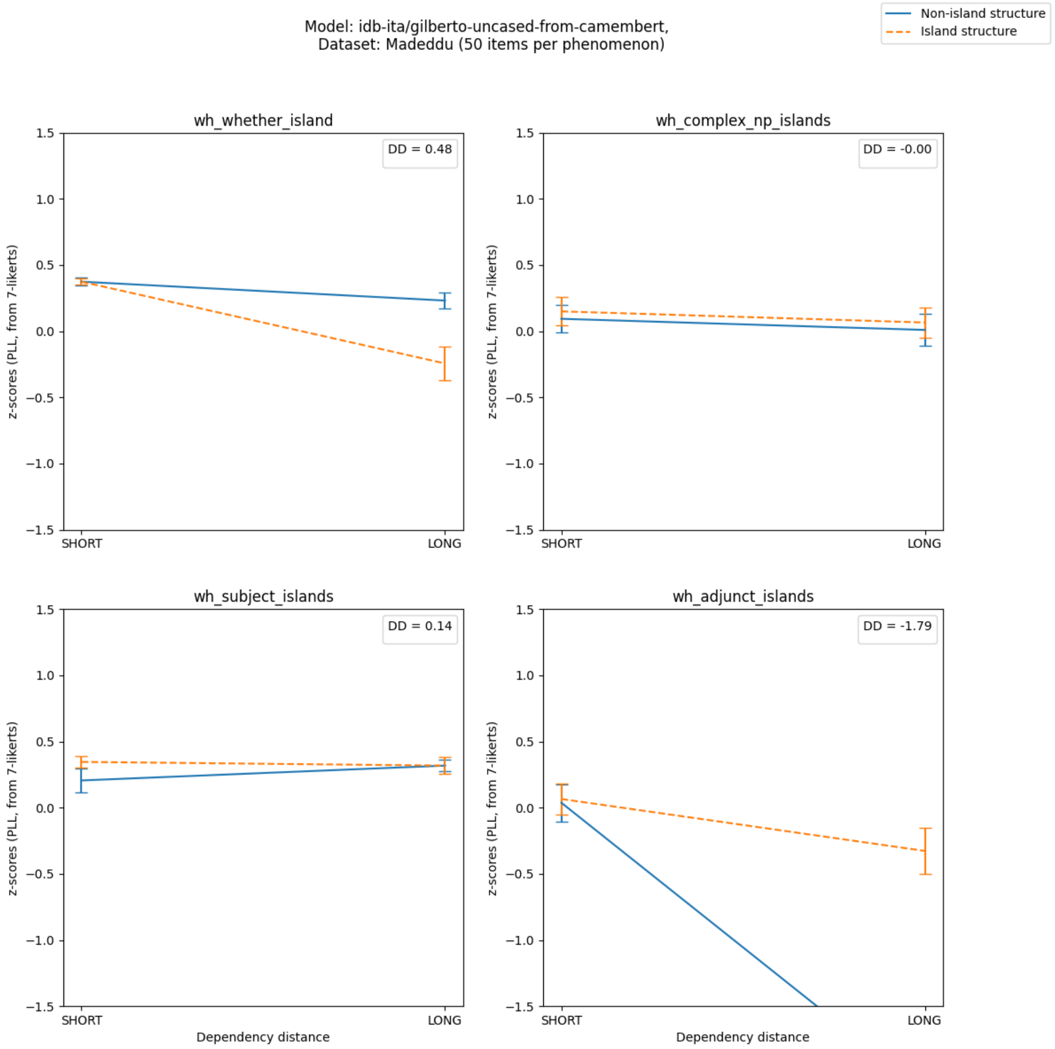
GilBERTo with PenLP (from softmax model output)



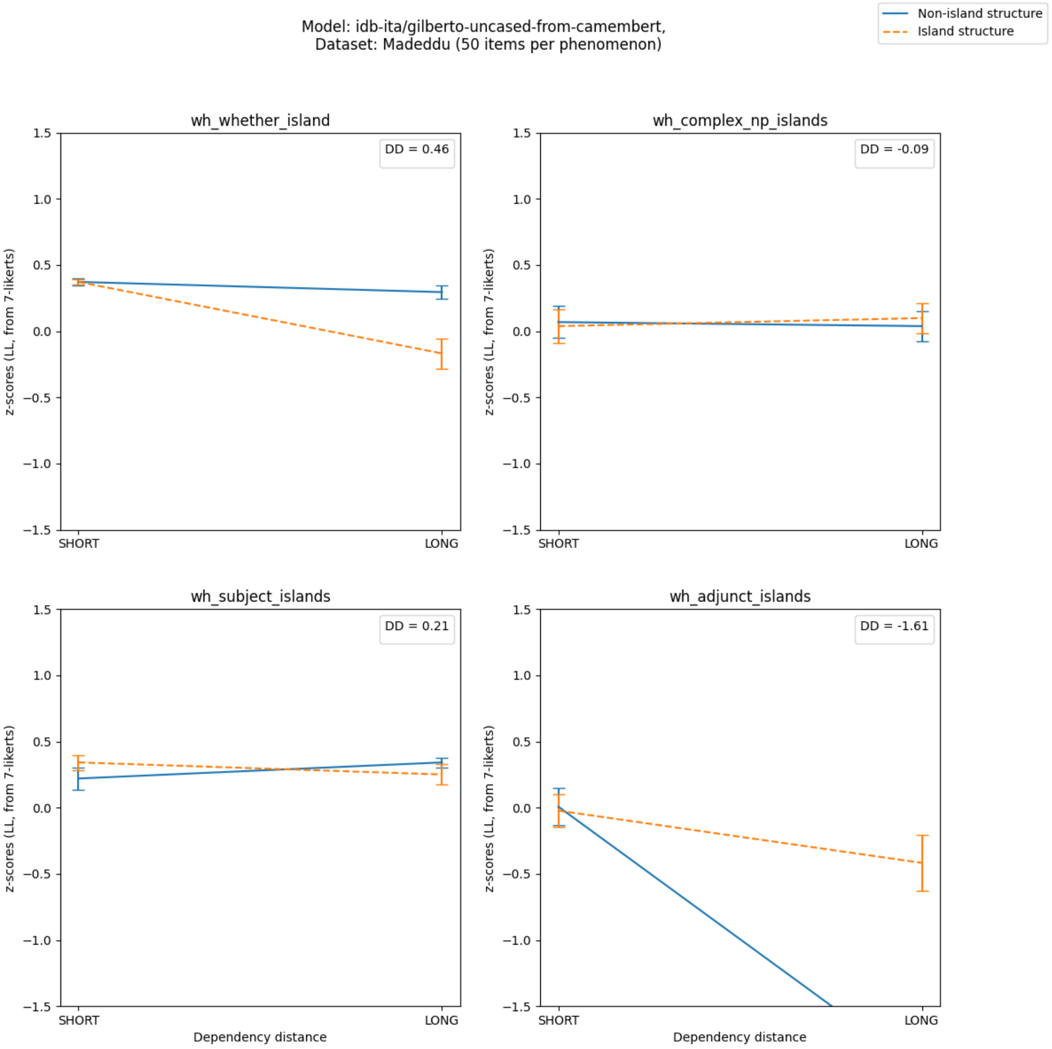
GilBERTo with LP (from softmax model output)



GilBERTo with PenLP-L (from logistic function model output)

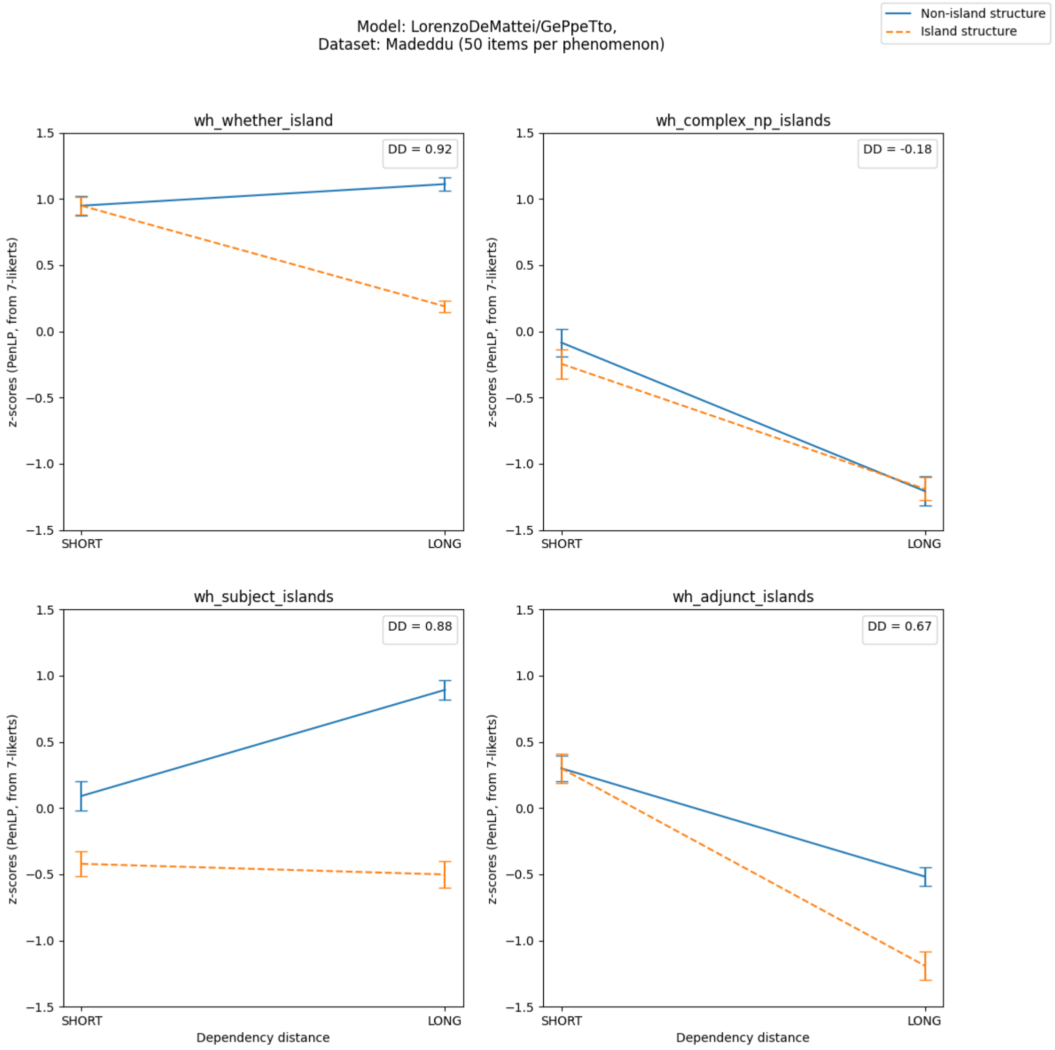


GilBERTo with LP-L (from logistic function model output)

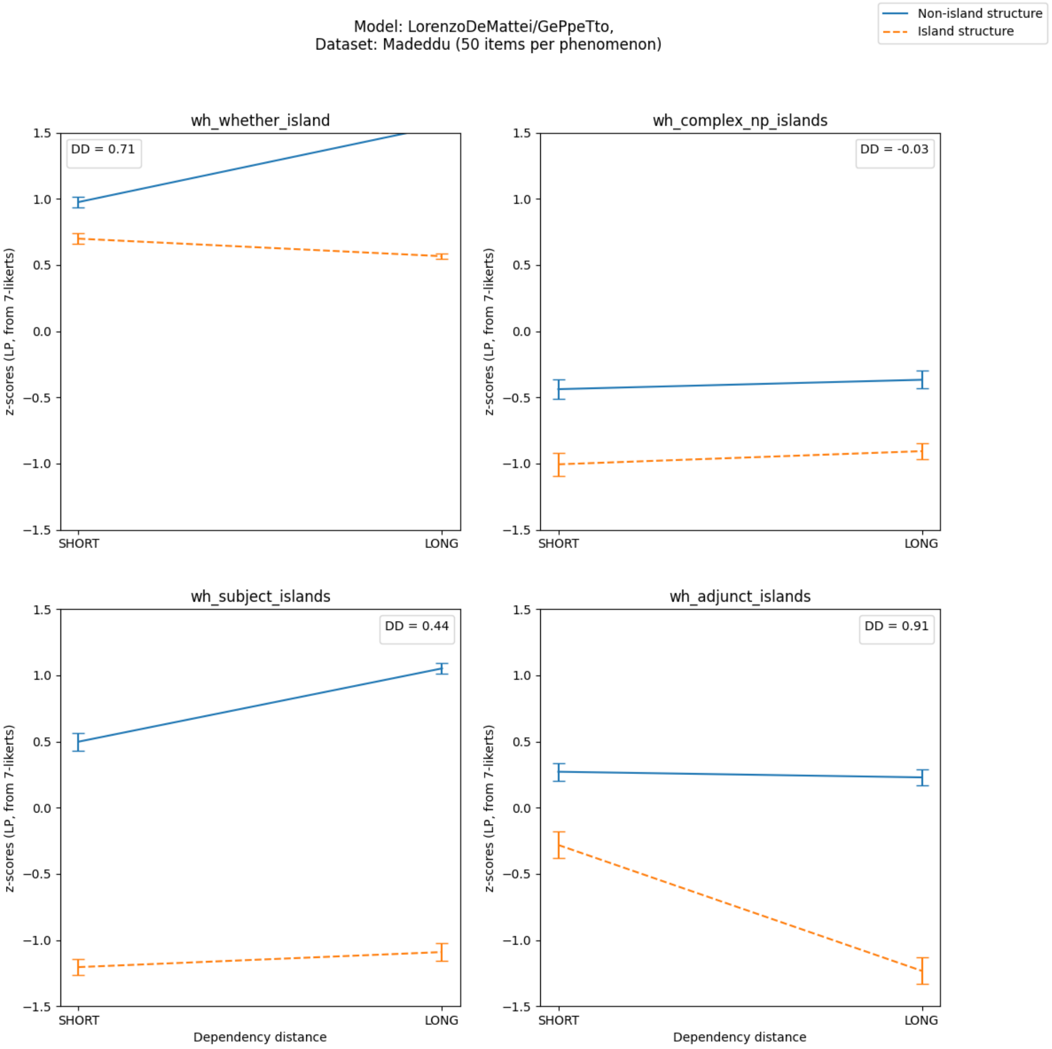


A.2.3 GePpeTto

GePpeTto with PenLP (from softmax model output)



GePpeTto with LP (from softmax model output)



..

Bibliography

Bostrom, K. and Durrett, G. (2020), Byte pair encoding is suboptimal for language model pretraining, *in* ‘Findings of the Association for Computational Linguistics: EMNLP 2020’, pp. 4617–4624.

Chaves, R. P. and Dery, J. E. (2014), Which subject islands will the acceptability of improve with repeated exposure, *in* ‘Proceedings of the 31st West Coast Conference on Formal Linguistics’, Citeseer, pp. 96–106.

Chowdhury, S. A. and Zamparelli, R. (2018), Rnn simulations of grammaticality judgments on long-distance dependencies, *in* ‘Proceedings of the 27th international conference on computational linguistics’, pp. 133–144.

De Mattei, L., Cafagna, M., Dell’Orletta, F., Malvina, N. and Guerini, M. (2020), Geppetto carves italian into a language model, *in* ‘Seventh Italian Conference on Computational Linguistics (CLIC-it 2020)’, Vol. 2769, p. 136.

Futrell, R., Wilcox, E., Morita, T. and Levy, R. (2018), ‘Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency’, *arXiv preprint arXiv:1809.01329* .

Hewitt, J. and 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)’, pp. 4129–4138.

Holtzman, A., Buys, J., Du, L., Forbes, M. and Choi, Y. (2019), The curious case of neural text degeneration, *in* ‘International Conference on Learning Representations’.

Hu, J., Gauthier, J., Qian, P., Wilcox, E. and Levy, R. P. (2020), A systematic assessment of syntactic generalization in neural language models, *in* ‘Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics’, pp. 1725–1744.

Jurafsky, D. and Martin, J. H. (2021), Speech and language processing (3rd ed. draft). Available from: https://web.stanford.edu/ jurafsky/slp3/.

Kush, D., Lohndal, T. and Sprouse, J. (2018), 'Investigating variation in island effects', *Natural language & linguistic theory* **36**(3), 743–779.

Lau, J. H., Armendariz, C., Lappin, S., Purver, M. and Shu, C. (2020), 'How furiously can colorless green ideas sleep? sentence acceptability in context', *Transactions of the Association for Computational Linguistics* **8**, 296–310.

Lau, J. H., Clark, A. and Lappin, S. (2017), 'Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge', *Cognitive science* **41**(5), 1202–1241.

Linzen, T., Dupoux, E. and Goldberg, Y. (2016), 'Assessing the ability of lstms to learn syntax-sensitive dependencies', *Transactions of the Association for Computational Linguistics* **4**, 521–535.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N. (2018), 'Generating wikipedia by summarizing long sequences', *arXiv preprint arXiv:1801.10198* .

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. and Levy, O. (2020), 'Emergent linguistic structure in artificial neural networks trained by self-supervision', *Proceedings of the National Academy of Sciences* **117**(48), 30046–30054.

Marvin, R. and Linzen, T. (2018), Targeted syntactic evaluation of language models, *in* 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', pp. 1192–1202.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018), Improving language understanding by generative pre-training, Technical report, OpenAI.

Ross, J. R. (1968), Constraints on Variables in Syntax, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Salazar, J., Liang, D., Nguyen, T. Q. and Kirchhoff, K. (2020), Masked language model scoring, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 2699–2712.

Sprouse, J., Caponigro, I., Greco, C. and Cecchetto, C. (2016), 'Experimental syntax and the variation of island effects in english and italian', *Natural Language & Linguistic Theory* **34**(1), 307–344.

Trotta, D., Guarasci, R., Leonardelli, E. and Tonelli, S. (2021), Monolingual and cross-lingual acceptability judgments with the italian cola corpus, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2021', pp. 2929–2940.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

von Prince, K. and Demberg, V. (2018), 'Pos tag perplexity as a measure of syntactic complexity'.

Wang, A. and Cho, K. (2019), Bert has a mouth, and it must speak: Bert as a markov random field language model, *in* 'Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation', pp. 30–36.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F. and Bowman, S. R. (2020), 'Blimp: The benchmark of linguistic minimal pairs for english', *Transactions of the Association for Computational Linguistics* **8**, 377–392.

Warstadt, A., Singh, A. and Bowman, S. (2019), 'Neural network acceptability judgments', *Transactions of the Association for Computational Linguistics* **7**, 625–641.

Wei, J., Garrette, D., Linzen, T. and Pavlick, E. (2021), Frequency effects on syntactic rule learning in transformers, *in* 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', pp. 932–948.

Wilcox, E., Levy, R., Morita, T. and Futrell, R. (2018), What do rnn language models learn about filler–gap dependencies?, *in* 'Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP', pp. 211–221.