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 intentionally left blank

Chapter 1

Introduction

1.1 Motivation

With the rapid progress made by deep neural language models in the last few years, during which they have significantly advanced the state of the art of many NLP benchmarks and applications, it has become increasingly important to understand what these model actually learn about human language (Rogers et al., 2020; Hewitt and Manning, 2019; Manning et al., 2020; Trotta et al., 2021).

Recent studies have shown that these models, without explicit supervision, automatically "rediscover" traditional linguistic knowledge (like morphological, syntactic, and semantic structures) (Rogers et al., 2020), partially mimicking the classical NPL pipeline (Tenney et al., 2019; Nikoulina et al., 2022; Wu et al., 2021). It has also been shown that that such knowledge is crucial for these models performance in downstream application tasks (Elazar et al., 2021).

However, the linguistic rediscovery that these model make has been found to differ from human-like generalizations (Linzen, 2020), and to be lacking in robustness, since they are prone to "frequency effects", in which the linguistic generalizations degrade when they need to be applied to less common words (Wei et al., 2021). Other works have shown that some of these models success in NLP Benchmarks is due to shallow heuristics, rather than being based on a general knowledge of language (Rogers et al., 2020).

A targeted linguistic evaluation of deep neural language models can detect and help to improve which of their linguistic generalizations are sub-optimal, which in turn could bring many benefits, including enabling hypothesis-driven improvement of the models architecture (Rogers et al., 2020), improving the performance on downstream application tasks, improving sample efficiency during training (Linzen, 2020), indicating more effective training strategies, and making more explainable the "black box" of these models internal representations (Wilcox et al., 2018). For these reasons, the evaluation of the

linguistic competence of these models has emerged as one of the core subfields in NLP (Cherniavskii et al., 2022). A recent survey (Rogers et al., 2020) listed the development of comprehensive stress tests for the different aspects of linguistic knowledge (which we still don't have) as arguably one of the most promising direction of research for modern language models.

This line of research is also of interest for addressing research questions in linguistics (Hewitt and Manning, 2019). For instance, investigating to what extent these self-supervised models, without any explicit prior bias, can acquire non-trivial syntactic skills to a human level (Gulordava et al., 2018; Warstadt et al., 2020), impacts the debate on the innateness of the language faculty (Hauser et al., 2002).

1.2 Our contributions and research questions

We contribute to the research on target syntactic evaluation of neural language models by assessing transformer-based language models trained in the Italian language on one of the most challenging syntactic phenomena, island effects.

We develop a test suite covering four types of syntactic islands in Italian: whether islands, adjunct islands, complex NP islands, and subject islands, all based on wh-dependencies and simple wh-words as fillers. The test suite is composed of 50 items per island type, each composed of 4 sentences, for a total of 200 stimuli.

We test both unidirectional (GPT-2) and bidirectional (BERT) transformer-based language models, all with a fixed size of 110M parameters, but trained on different corpora of different sizes, from 2B to 13B tokens.

For our methodology, we apply a factorial experimental setup adapted from psycholinguistic research, subtracting out the influence of multiple effects and controlling confounds. We compare our results with a previous study (Sprouse et al., 2016) on human subjects using a similar format for the same island effects in Italian. Following Hu et al. (2020), we also define a factorial success criterion to obtain a final accuracy score, and compare our results with previous work testing neural language models on island effects in the English language. We then try to explain the results by making a further analysis of the per-token surpisal scores of bidirectional models on some items. We find that this type of analysis to the individual token scores is very effective in explaining the results and what linguistic factors seem to influence more the models responses.

To our knowledge, this is the first work on the targeted syntactic evaluation of Italian transformer-based models that uses a factorial design and covers island effects.

The research questions we try to address with the present work are the following:

Do the models responses to island effects resemble those from humans? Do their acceptability judgements seem to be affected by the same island effects in the same way? Are the accuracy scores of Italian language models on island effects comparable to those of English models on the same syntactic island types?

Do we see any correlation between the model performance and the amount and quality of training data, or their architecture (unidirectional vs bidirectional)?

Is our methodology effective in assessing language models on island effects? Which factors should be controlled the most for confounds?

1.3 Strengths and limitations of transformer-based language models

Transformer-based language models were introduced in 2017 (Vaswani et al., 2017), and since they they have revolutionized the NLP field, substantially advancing the state of the art in many benchmark and applications.

These models have been shown to be able to encode non-trivial hierarchical linguistic representations, for instance akin to syntactic tree structures (Rogers et al., 2020). Other linguistic knowledge that has been shown that can be extracted from these models include word order information, part of speech, syntactic chuncks, and semantic roles (Tenney et al., 2019; Liu, Gardner, Belinkov, Peters and Smith, 2019; Rogers et al., 2020). In some cases this knowledge has been extracted using probing classifiers, although this methodology is debated because it is not always clear if part of the knowledge was actually learned by the model during the fine-tuning for the probing classifier (Rogers et al., 2020).

The impressive advancements made by these models are however countered by some of their limitations, that pose doubts about what these models actually understand about language. BERT has been shown to struggle or be unable to acquire linguistic skills that require pragmatic knowledge, common sense, and basic reasoning. For instance, BERT is not able to understand negation, to make basic arithmetic operations on numbers (Wallace et al., 2019), to use pragmatic inference and role-based event knowledge (Ettinger, 2020), and to grasp the abstract attributes of objects (Da and Kasai, 2019). Ettinger (2020) found that "BERT struggles with challenging commonsense/pragmatic inferences and role-based event prediction; that it is generally robust on within-category distinctions and role reversals, but with lower sensitivity than humans; and that it is very strong at associating nouns with hypernyms. Most strikingly, however, we find that BERT fails completely to show generalizable understanding of negation, raising questions about the aptitude of LMs to learn this type of meaning". Undoubtedly, some of these

limitations are due to the fact BERT is trained in the single-modality of raw text, without any grounding in vision or other perceptions.

Reasoning about the knowledge it possess seem to be the hardest task overall for BERT. While it can guess affordances and properties of many objects (e.g. the fact that people can walk into houses, and that houses are large), it cannot infer the relationship between their properties (e.g., the fact that houses are larger than people) (Forbes et al., 2019).

These limitations are reflected also in the massive amount of data needed to train these models. While BERT seems to be able to learn most syntax and semantics with about 100M of pretraining tokens, to learn Common sense knowledge and reasoning skills over 30B tokens are needed (Zhang et al., 2021), which is several orders of magnitude more than the amount needed by humans to learn language (Linzen, 2020). In addition to this sample inefficiency, BERT seems to also under-utilize its parameters, since many of them and its attention heads have been shown that can be pruned without a significant loss in performance (Rogers et al., 2020).

1.4 Targeted linguistic evaluation of language models

Determining which linguistic phenomena are actually learned by modern transformer-based language models has become increasingly important, since they are trained without any explicit linguistic supervision, and their internal representations are akin to "black boxed" very hard to interpret. Investigating the linguistic competence acquired by these models serves multiple purposes, like having a clear picture of their strengths and limitations, to identify areas for improving their robustness and their ability to generalize in a human-like way, and to accelerate work towards building general-purpose models for language understanding (Wilcox et al., 2018; Ettinger, 2020).

There has been a growing effort in this direction of research, which has been dominated in particular by the assessment of these models syntactic skills (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Wilcox et al., 2018; Chowdhury and Zamparelli, 2018; Futrell et al., 2019; Ettinger, 2020).

This relatively new area of research has met methodological problems in the non trivial task of assessing these models abilities and internal representations. The methodology has not yet converged to an established paradigm, multiple approaches are currently in use, including adapting experimental procedures from psycholinguistic (Futrell et al., 2019). Developing new, more effective and more comprehensive methodologies for assessing the linguistic competence of modern language models is still an open area of research (Ribeiro et al., 2020; Newman et al., 2021).

One of the main approaches currently in use is that of targeted syntactic evaluations, which uses minimal pairs of sentences, of which the model is expected to give the higher probability to the grammatical sentence rather than the ungrammatical ones. These minimal pairs are supposed to vary only over a single condition, in order to isolate the ability of the model in a particular phenomenon and control confounds. However, building these minimal pairs can be labor-intensive, and it is not always easy to design minimally varying sentences for some more complex phenomena (Warstadt et al., 2020). Additionally, recent studies have found that some current targeted linguistic evaluations approaches might be overestimating these models linguistic skills (Newman et al., 2021).

For more complex syntactic phenomena, like Negative Polarity Items (NPIs) or constraints on Filler-Gap dependencies, other approaches have been started to be utilized, like a factorial experimental setup, imported from the methodology in psycholinguistic research (Wilcox et al., 2018). A factorial experimental setup can be considered as a generalization of a minimal pair ones, in which and item has more then two sentence, all varying minimally covering the all the combinations of particular factors or property (for instance, a 2x2 factorial design is common). This allows to quantify and factor out confounding factors, by calculating the difference in score between sentences exemplifying different conditions.

1.5 On transformer-based language models

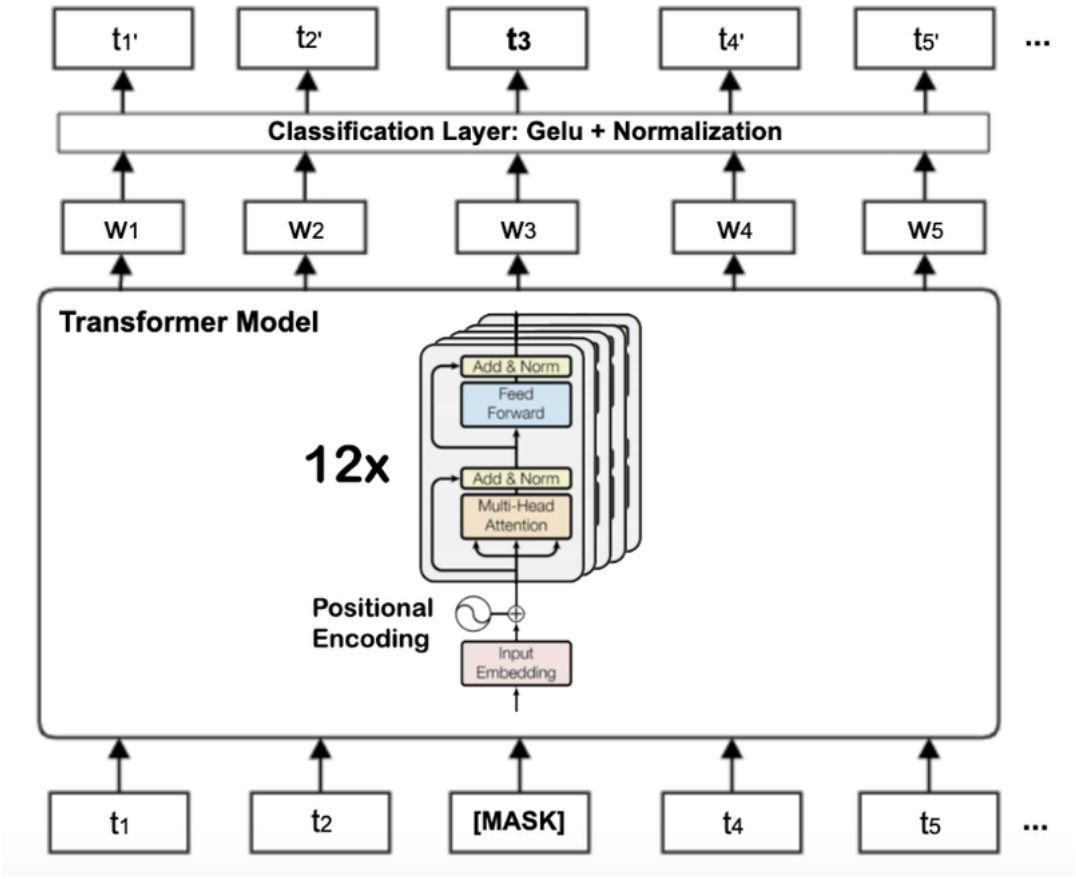
The Transformer neural network architecture (Vaswani et al., 2017) takes a sequence of tokens as input, which processes through a series of layers of "transformer blocks", and produces in output a contextualized representation of each token, in the form of vector embeddings (Manning et al., 2020).

The transformer blocks are based on the "attention mechanism", which is a form of generalized dot-product that models all pairwise interactions between each token and any other token in the input sequence.

For each input sequence, an attention "head" assigns a weight to each pair of tokens, indicating how much "attention" the model should pay to the first word when modeling properties of the second one. Each transformer block can have multiple attention heads in parallel, each capturing a different token–token dependency relationships between words. For each token, the information from each head in a layer block is aggregated into a vector embedding, which is a contextualized representation of that token in the input sequence (Manning et al., 2020).

A typical task to train these model is the Masked Language Model objective, or cloze task, in which a portion of each sequence is masked (typically 15% of the tokens) and the model must predict the masked words given the unmasked ones. The probability that the model assigns to each masked word

is used to calculate the cross-entropy loss for back-propagation.



(Image taken from Khalid et al. (2021), license CC BY 4.0)

Figure 1.1: The Transformer based BERT base architecture with twelve encoder blocks.

Transformer blocks, unlike recurrent neural networks (RNNs), have the computational advantage of not being recurrent, and therefore are highly parallelizable, faster to compute, and scale in a way that RNNs cannot (Lau et al., 2020). Transformer based language models contain several millions of trainable parameters. A common baseline model for BERT is composed of 110 million parameters spread over 12 layers of transformer blocks (Figure 1.1).

They typically require massive amount of data (in the order of at least hundreds of million tokens) and computation time to achieve state of the art performance in natural language processing tasks, and are therefore very expensive to train from scratch, although they can outperform the previous generation of neural language models when properly optimized (Manning et al., 2020).

Chapter 2

A Brief Introduction to Island Effects

An *island effect* is the phenomenon of a sentence becoming ungrammatical or unacceptable when there is a long-distance "filler-gap" dependency which crosses the boundaries of certain constructs (as in (4)), which have been metaphorically called syntactic *islands* (Ross, 1967; Sprouse and Hornstein, 2013).

In psycholinguistic terminology (since (Fodor, 1978)) a filler–gap dependency is the relationship between the antecedent or "filler" (e.g. a complementizer like "*what*" or "*who*" in English, or "*cosa*" or "*chi*" in Italian) and the canonical position (often called "gap", as a postulated empty position or element with no surface manifestation) of the predicate argument that it replaces (Pearl and Sprouse, 2013; Wilcox et al., 2018; Hawkins, 1999). Other terms to refer to filler–gap or long-distance dependencies are syntactic movement or *extraction*[1] (Jurafsky and Martin, 2021).

Consider the following examples (taken from Sprouse et al. (2016)), in which (1) is a plain sentence with no "extraction", and (2) contains a filler-gap dependency (the gaps are marked with underscores, and the fillers in bold) :

(1) Susan thinks that John bought the book.

(2) **What** does Susan think that John bought _?

The dependency in (2) is called a wh-interrogative clause dependency (wh-dependency, in short) because the filler is an interrogative wh-word[2]. There

[1]The use of the term *extraction*, to refer to filler-gap dependencies, evokes the idea of movement formulated in transformational grammars, within the generative linguistic paradigm, as if the filler is moved or extracted from the canonical argument position (Sprouse and Hornstein, 2013).

[2]More generally, it can be a simple wh-phrase, like "*who*" or "*what*", or a complex wh-phrase, like "*which book*."

are however other types of filler-gap dependencies that are blocked by syntactic islands, like relative clause dependencies (rc-dependencies, in short)[3]. Island effects have been shown to display different patterns of variations with different dependency types, but in the present work we'll focus only on wh-dependencies.

In (1) the argument *the book* is adjacent to the predicate that selects it (*bought*), while in (2) the wh-word *What* is not. Because of this, syntacticians sometimes call wh-dependencies as in (2) *long-distance* dependencies, distinguishing them from *short-distance* dependencies as in (1) where the two constituents in the dependency relationship are adjacent or nearly adjacent to each other (Pearl and Sprouse, 2013).

The distance between the filler and the gap can be made arbitrarily long (as long as human working memory capacity permits), unless a syntactic island intervenes, as in the following examples (taken from Sprouse and Hornstein (2013))[4]:

(3) What does Susan think that Lily said that Sarah heard that John bought _?

(4) Whether island:
*What does Susan wonder [whether John bought _]?

It is worth mentioning here that island effects have been found to decrease with repeated exposure, that is, an individual repeatedly exposed to an island structure such as the one in (4), will find the resulting sentences acceptable, or less unacceptable (Chaves and Dery, 2014).

Island effects are usually named after the type of syntactic structure that creates them (Sprouse and Hornstein, 2013). The sentence in (4) exemplifies an extraction from a wh-island, specifically of the whether-island sub-type, that is from an embedded clause headed with the complementizer "whether" (Wilcox et al., 2018). Whether islands are among the most easily intelligible types of islands for learners (Pearl and Sprouse, 2013), along with Adjunct islands (5), Complex Noun Phrase islands (6) and Subject islands (7), which are the four types on which we'll focus in the present thesis. The following examples are taken from Sprouse et al. (2016):

(5) Adjunct island
*What does Susan worry [if John buys _]?

[3]In rc-dependencies, the filler can be either a relative pronoun or a head noun, and it introduces a headed relative clause that hosts the gap, as in the example (from Sprouse et al. (2016)) "*I like* ***the car*** *that you think [that John bought _].*"

[4]In the examples we use the following notation conventions: the asterisks indicate unacceptable sentences, as it is common in linguistic literature. The syntactic islands constructs, and occasionally other clauses, are marked within square brackets, but this is for explanatory purposes only, they are not part of the sentences given as input to the language models. The gaps are marked with underscores, again for explanatory purposes only.

(6) Complex NP Island
*What did Susan make [the claim that John bought _]?

(7) Subject island
*What did Susan think that [the speech about _] interrupted the TV show?

In the adjunct island example above (5), there is the extraction of a constituent (the direct object) from an embedded sentential adjunct clause (an adjunct clause is an optional clause which can be omitted without affecting the grammaticality of the main clause (Downing and Locke, 2002)).

A complex noun phrase is a phrase headed by a noun ("the claim" in (6)) with a modifier that is another phrase ("that John bought"). In (6), there is an extraction of a direct object argument (the thing that is bought by John) from the subordinate complex noun phrase ("[the claim that John bought _").

In (7), there is an extraction from a propositional phrase ("about _") that modifies a noun phrase ("[the speech about _]") that occurs in the subject position (for the verb "interrupted") of a subordinate clause (Wilcox et al., 2018; Sprouse et al., 2016).

2.1 The importance of island effects in linguistics

The linguistic research on island effects has received prominence for their role in arguments on the innate nature of the language faculty. A staple of proponents of the linguistic paradigm known as *generative linguistics*, which since its first formulation by Chomsky (1957) has been the hegemonic linguistics paradigm for decades, is the Universal Grammar hypothesis (UG) (Chomsky, 1965), according to which the language faculty is innate, and the acquisition of language is made possible by the presence of innate linguistic constraints or biases in the human brain (Dąbrowska, 2015). According to the most common definitions, the "universal grammar" present in the brain would consist of an innate "system of categories, mechanisms and constraints shared by all human languages", which includes "principles", general constraints to which all human language grammars must abide, and "parameters", which are the possible variations in grammatical features between languages (Dąbrowska, 2015).

One of the main and most influential argument in support of the Universal Grammar hypothesis, has been the Argument from the Poverty of the Stimulus (APS), according to which linguistic knowledge possessed by children cannot be acquired exclusively from the input which is available to them, therefore they must possess some innate knowledge of it (Pullum and Scholz, 2002).

Island constraints have been central in proposals for the innate presence of complex linguistic biases in the human brain, for the Argument from the Poverty of the Stimulus, and in the debates on generative UG-based syntactic theories (Pearl and Sprouse, 2013). This is due to the fact that island effects are a complex phenomenon, rare or non-existent in child-directed speech (and of relative infrequency even in adult-directed speech), which therefore "raise difficult questions about how children could use their limited input to arrive at a grammar that includes long-distance dependencies that are nonetheless constrained by specific structural configurations. In this way, island effects provide a classic motivation for theories that assume domain-specific constraints on language acquisition (i.e. universal grammar)" (Sprouse et al., 2012).

For this reason, since their discovery by Ross (1967), they have been heavily investigated and been a subject of debate in the linguistic and psycholinguistic literature, with hundreds of articles in dozens of languages devoted to their investigation (Sprouse and Hornstein, 2013; Pearl and Sprouse, 2013).

One of the main arguments against the existence of proper island "constraints" that render a sentence ungrammatical when violated, is that such ungrammaticality or unacceptability could arise instead discourse-structural factors or simply from the increased *processing difficulty* due to the sum of other factors present in alleged "island violations", like the fact of having a long-distance dependency combined with more marginal syntactic constructs (Sprouse and Hornstein, 2013; Wilcox et al., 2018; Ambridge and Goldberg, 2008; Hofmeister and Sag, 2010).

To address poverty of the stimulus arguments regarding island effects, a psycholinguistic study by Pearl and Sprouse (Sprouse and Hornstein, 2013) implemented a computational model able to learn island constraints corpus of child-directed speech, without any prior linguistic bias (like a Universal Grammar). Modern deep neural language models have also began to be used as an argument that a system without any prior linguistic learning biases can learn non-trivial syntactic phenomena and other hierarchical structures (Linzen, 2018). For instance Gulordava et al. (2018) wrote: "We tentatively conclude that LM trained RNNs can construct abstract grammatical representations of their input. This, in turn, suggests that the input itself contains enough information to trigger some form of syntactic learning in a system, such as an RNN, that does not contain an explicit prior bias in favor of syntactic structures". And (Warstadt et al., 2020): "By evaluating whether self-supervised learners like LMs acquire human-like grammatical acuity in a particular domain, we gather indirect evidence as to whether this phenomenon is a necessary component of humans' innate knowledge".

2.2 Cross-linguistic variation in island effects

Cross-linguistic variations have been observed for island effects, since the first study on island effects in Italian by Rizzi (1982). While English shows at least eight different types of island effects (whether islands, complex NP islands, subject islands, adjunct islands, relative-clause islands, sentential subject islands, coordinate structure constraint violations, and left branch extraction violations), Scandinavian languages (Swedish, Norwegian, Danish, and Icelandic), show almost no island effects[5] (Pearl and Sprouse, 2013). Here are examples of the above mentioned eight types of syntactic islands in English, taken from Sprouse and Hornstein (2013):

Example 2.2.1. Island effects in English

a. WHETHER ISLAND:
*What do you wonder [whether John bought _]?

b. COMPLEX NP ISLAND:
*What did you make [the claim that John bought _]?

c. SUBJECT ISLAND:
*What do you think [the speech about _] interrupted the TV show?

d. ADJUNCT ISLAND:
*What do you worry [if John buys _]?

e. RELATIVE-CLAUSE ISLAND:
*What did you meet [the scientist who invented _]?

f. SENTENTIAL SUBJECT ISLAND:
*What did [that John wrote _] offend the editor?

g. COORDINATE STRUCTURE CONSTRAINT VIOLATION:
*What did John buy [a shirt and _]?

h. LEFT BRANCH EXTRACTION VIOLATION:
*Which did John borrow [_ book]?

Most island effects are present in Romance languages, with some exceptions, as some types of syntactic islands have effect in certain types of sentences, but not in others. This is the case of subject islands, which are not

[5] Sprouse and Hornstein (2013) report that "Swedish is not bereft of apparent island effects. Rather it does not display island effects in all contexts where they are theoretically expected to appear (and as they do appear in English). For example, there are some unacceptable instances of extracting out of complex noun phrases, but others seem perfectly fine."

present in Italian, Spanish and Portuguese, while they are present in French (Pearl and Sprouse, 2013). In a psycholinguistic study, Sprouse et al. (2016) found evidence that subject islands with relative clause dependencies are not present in Italian.

There is also a substantial varying degree of acceptability between the different types of islands. For instance, wh-islands are generally among the strongest sources of sentence acceptability, while violations of complex NP islands with rc-dependencies tend to be more acceptabile (Pearl and Sprouse, 2013). Because of this, it has been proposed in the literature a distinction between strong and weak islands (Sprouse et al., 2016; Szabolcsi, 2006). Furthermore, the acceptability of island violations also varies from speakers to speaker within the same language, with some speakers being more sensitive than others to island effects (Sprouse and Hornstein, 2013).

2.3 Island effects in Italian

Island effects in Italian were first studied by Rizzi (1982), which observed some cross-linguistic variation between English and Italian. While most studies of island effects in English usually focus on syntactic island built with wh-interrogatives, in his study Rizzi focused only on syntactic islands under relative clause dependencies, arguing that it was problematic study them with wh-dependencies because they would require a double wh-word (as in **Chi ti domandi chi _ ha incontrato _?*)(*Who do you wonder who met?*), which itself renders a sentence unacceptable, resulting in a confound. However, Sprouse et al. (2016), which assessed the same types of islands with wh-dependencies in English and Italian speakers, showed that is possible to circumvent this problem, by studying wh-dependencies with island types that do not require a second wh-word. Specifically, for a direct cross-linguistic comparison, they studied wh-dependencies with whether-islands in English, and *se*-islands ("*if*") in Italian.

Here are some of the examples of islands with rc-dependencies studied by Rizzi (1982) (the translations are from Sprouse et al. (2016)):

(8) Wh-island
Tuo fratello, a cui mi domando che storie abbiano raccontato _ _, era molto preoccupato.
"Your brother, who I wonder what stories they told, was really worried."

(9) Complex NP Island
*Questo incarico, che non sapevo [la novità che avrebbero affidato _ a te], ...
"This task, which I didn't know the new that they may have assigned

to you..."

(10) SUBJECT ISLAND
Questo autore, di cui so che [il primo libro _] è stato pubblicato recentemente, ...
"This task, which I didn't know the new that they may have assigned to you..."

In the present thesis, we focus on assessing island effects under wh-dependencies, which have been shown to produce island effects in both English and Italian (Rizzi, 1982; Sprouse et al., 2016). The following (11-14) are examples in Italian with wh-dependencies and the four island types administered to language models in our experiments:

(11) WHETHER ISLAND
*Cosa ti domandi [se io abbia comprato _]?
What yourself wonder if I have bought
'What do you wonder if I bought?'

(12) ADJUNCT ISLAND
*Che cosa Gianni è partito per Parigi [dopo aver fatto _]?
What Gianni is left for Paris after having done
'What did Gianni leave for Paris after packing?'

(13) COMPLEX NP ISLAND
*Cosa hai fatto [l'affermazione che il tuo amico avrebbe
What have done the statement that the your friend have
mandato _]?
sent
'What did you make the claim that your friend sent?'

(14) SUBJECT ISLAND
*Di chi pensi che [la moto _] abbia urtato l'auto di
of who think that the motorbike have hit the car of
Chiara?
Chiara
'Who do you think the motorike of hit Chiara's car?'

In all of the items, the fillers are simple wh-words (e.g. "*cosa*", "*chi*")("*what*", "*who*"), rather than complex wh-phrases ("*che libro*")("*which book*") This is because these are the most commonly studied types of wh-dependencies for syntactic islands, and also to match the examples is (Sprouse et al., 2016) for comparison purposes. The type of wh-words (simple vs complex) is one of

the factor that in principle could affect the processing of island effects and therefore their acceptability.

One aspect that influences the difference in sentence structure of syntactic islands in Italian, is the fact that Italian does not allow *preposition stranding as in English*. Preposition stranding is the possibility of having a preposition *stranded* (separated from) its object. For example, in the English sentence *I know **who** you bought the painting **from***, "from" and "who" are stranded. Instead, Italian forces "*pied-piping*" (another metaphorical technical term introduced by Ross (1967), like that of "islands", as mentioned in chapter 2), the fact that the prepositional phrase must be continuous (the wh-word *brings along* the preposition): *So **da chi** hai comprato il quadro* ("*I know from who you bought the painting*").

2.4 Factorial experimental setups from psycholinguistic studies

A factorial experimental setup allows to quantify and tease out the effect of different factors that contribute to a phenomenon, in order to better control confounds (Sprouse et al., 2016, 2012). Compared to a minimal pairs approach to language models assessment, a factorial test design is better suited to assess more complex syntactic phenomena like island effects (Warstadt et al., 2020).

A common use case is to have ITEMS (the equivalent of the sentence pairs in a minimal pairs approach) structured in a 2×2 paradigm, with four sentences in total, each across two binary PROPERTIES. Each sentence exemplifies a CONDITION, which is a particular state of the two properties. For instance, in the case of (Wilcox et al., 2018) factorial assessment of filler-gap dependencies, the two properties were the presence or absence of a filler, and the presence or absence of a gap. The same approach can be generalized to more complex paradigms, with more than two properties, each having more than two LEVELS (as is the case for binary properties).

Consider the item in the following example (15), in which condition a. displays a short-distance dependency and no island structure (SD/NI); condition b. displays a long-distance dependency and no island structure (LD/NI); condition c. displays a short-distance dependency and an adjunct island structure (SD/IS); and condition d. displays a long-distance dependency crossing an adjunct island structure (LD/IS). Sentences a-c are supposed to be acceptable, while d is unacceptable because of an island effect:

(15) ADJUNCT ISLAND

a. Chi dice che l'imprenditore ha raccolto i fondi
Who say that the businessman have raised the funds
sufficienti? [SD/NI]
sufficient

'Who says that the businessman raised sufficient funds?'

b. Che cosa dici che l'imprenditore ha raccolto _? [LD/NI]
What say that the businessman have raised
'What do you say that the businessman raised?'

c. Chi ha avviato l'attività [quando ha raccolto i fondi
Who have started the business when have raised the funds
sufficienti]? [SD/IS]
sufficient
'Who started the business when he raised sufficient funds?'

d. *Che cosa l'imprenditore ha avviato l'attività [quando
What the businessman have started the business when
ha raccolto _]? [LD/IS]
have raised
'What did the businessman started the business when he raised?'

In a factorial design, the scores (being either a human judgment on a likert scale, or a language model score like a log probability or a pseudo-log-likelihood)given to the sentences of an item like (15), can be compared to quantify and tease out the effect of different factors. For instance, in the case of a pseudo-log-likelihood (PLL) score obtained from a bi-directional LM like BERT, we could calculate the following quantities:

- LENGTH EFFECT $= PLL_a - PLL_b$
- STRUCTURE EFFECT $= PLL_a - PLL_c$
- TOTAL EFFECT $= PLL_a - PLL_d$
- ISLAND EFFECT = TOTAL EFFECT$-$(LENGTH EFFECT+STRUCTURE EFFECT)

The LENGTH EFFECT should capture the decrease in sentence acceptability due to the increase of the dependency distance from short to a long filler-gap dependency. The STRUCTURE EFFECT should capture the effect of adding an island structure. The TOTAL EFFECT should capture the whole decrease in sentence acceptability due to the increase in sentence complexity from the two factors and also due to the violation of the island constraint. Subtracting the first two effects from the total effect should isolate the ISLAND EFFECT alone, without considering the sentence acceptability decrease due to the other two factors. We conclude that a model has 'learned' the island effect displayed in an item, if the island effect score on that item is greater than zero.

This factorial score is used by Sprouse et al. (2016) to assess island effects as perceived by human subjects, and it is equivalent to a differences-in-differences (DD) score (Maxwell and Delaney, 2003). Similar factorial scores have been used for language models by Wilcox et al. (2018) and Hu et al. (2020).

The factorial design allows to calculate the effect of the combination of the different properties, by making subtractions between the scores of the sentences of an item, because they differ minimally between each other (only on the basis of the controlled properties). This allows to factor as many confound as possible, even unexpected ones, as long as they are evenly distributed within across the sentences of an item (Sprouse et al., 2016; Hu et al., 2020).

As Pearl and Sprouse (2013) explain "translating each of these properties into separate factors, each with two levels (dependency GAP POSITION: matrix, embedded; STRUCTURE present in question: non-island, island)," allows "to define island effects as a superadditive interaction of the two factors (..) in other words, an island effect is the additional unacceptability that arises when the two factors are combined, above and beyond the independent contribution of each factor. Specifically, a syntactic island occurs when there is more unacceptability than what the EMBEDDED dependency and the presence of an ISLAND structure in the question contribute by themselves" (Pearl and Sprouse, 2013). "An acceptability judgment experiment that employs a factorial definition of island effects. First, we can isolate the effect of dependency length on acceptability by contrasting a sentence with a short WH-dependency, an extraction from a matrix clause, with a sentence that contains a longer WH-dependency, an extraction from an embedded clause. Similarly, we can isolate the effect of processing island structures by contrasting a sentence with an island structure with a sentence that does not contain an island structure. Finally, we can measure the effect on acceptability of processing both long-distance WH-dependencies and island structures—the island effect itself—by combining both in a single sentence" (Sprouse et al., 2012).

Chapter 3

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 assess island effects or filler-gap dependencies, on neural language models in the case of Wilcox et al. (2018); Hu et al. (2020), and on human subjects in the case of Sprouse et al. (2016)

Hu et al. (2020) used a factorial design in which each test items is composed of 4 minimal varying sentences, and obtains an overall percentage accuracy score on a test suite using a success criteria for items. An item is considered to have been scored accurately by the models when multiple conditions are met (e.g. 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) obtain a measure of statistical significance and a confidence interval, instead of an accuracy score. In the case of Sprouse, however, the format of sentences employed shows more variation across the sentences of each item than those typically used for scoring language models, where usually all the sentences in an item are minimally different in terms of lexical content, and differ in just one word. Wilcox et al. (2018) circumvents the need for minimally different sentences, by scoring separately items with different constructs (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, and not on bi-directional language models like BERT.

3.1 Approaches for targeted linguistic evaluation of language models

3.1.1 Grammaticality vs Probability

While linguists are usually concerned with grammaticality or *acceptability* judgments when assessing linguistic expressions, computational language processing has traditionally been more concerned with *likelihood*, i.e. the probability of a sentence being produced or encountered (Lau et al., 2020). But the concepts of grammaticality and probability, while related and partially overlapping, do not coincide. In principle, a sentence can appear with very low probability but still be grammatically well-formed, such as the known example from generativist linguistics *Colorless green ideas sleep furiously* (Hu et al., 2020). In other words, while computational language models learn a probability distribution of word sequences, this does not corresponds exactly to a grammaticality or acceptability judgment (Marvin and Linzen, 2018).

There is still no generally accepted method for obtaining binary acceptability predictions from unsupervised language models, and this remains a fundamental research question (Lau et al., 2020; Warstadt et al., 2020)[1].

The concept of acceptability is broader than that of grammaticality. Grammaticality usually refers to syntactic acceptability (well-formedness), but acceptability in general includes also semantic, pragmatic, and even non-linguistic factors, like sentence length (Lau et al., 2020). Additionally, acceptability is usually a graded judgment, while grammaticality is a binary one (a sentence can be either grammatical or not).

Before the introduction of methods for extracting acceptability judgments from neural language models, the primary performance metric for their evaluation was **perplexity**, which is a broad-coverage metric that scores how well, on average, a model predicts a word in its contex, and captures how well the probability distribution learned by the model conforms to that of a particular linguistic domain (Marvin and Linzen, 2018; Hu et al., 2020).

As Chowdhury and Zamparelli (2018) explain, "Perplexity measures how many equally probable words can follow a point in the text; as the sentence grows longer and more information accumulates, the options for the following word decrease." It is "based on the intuition that an ungrammatical sentence should 'confuse' the NN more than a corresponding grammatical one, and that this confusion will translate in a decreased ability to make correct predictions." (Chowdhury and Zamparelli, 2018)

However, perplexity cannot give detailed insight into these models' knowledge of grammar, as it has been found to be at least partially dissociable from

[1] On a related note, even in the psycholinguistic literature and in theoretical linguistics there is no generally established metric for graded grammaticality judgments (Chowdhury and Zamparelli, 2018; Cowart, 1997; Sorace and Keller, 2005).

scores capturing specifically the grammatical skills of language models (Marvin and Linzen, 2018; Hu et al., 2020; Warstadt et al., 2020). Perplexity is not suitable to estimate ungrammaticality from a language model, since it does not locate the source of ungrammaticality at a specific point in a sentence (Chowdhury and Zamparelli, 2018). As noted by Marvin and Linzen (2018) "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."

While perplexity is not suitable as an absolute measure of a sentence grammaticality, it can be used as a relative measure between a minimal pair of sentences, varying only for a property that makes one sentence grammatical and the other ungrammatical, as has been done with the targeted syntactic evaluation (TSE) assessment paradigm. The introduction of the targeted linguistic evaluation of language models, and the minimal pairs paradigm in particular, therefore serves also as a supplement to perplexity as a complementary measure that capture other aspects of a model performance. Additionally, fine-grained linguistic assessment can provide more challenging evaluations, and it is also a tool that help in the explainability of these models, since the test results from fine-grained linguistic assessment are more interpretable (Hu et al., 2020).

The targeted linguistic evaluation of deep neural language models trained on raw textual input started with the work of Linzen et al. (2016), which used a corpus of naturally occurring sentences, extracted from the English Wikipedia, and tested the ability of Recurrent Neural Network models (RNNs), and in particular models based on long short-term memory units (LSTM). Linzen et al. used different testing approaches that will be developed in future work: (1) fine-tuning the models on a probing subject-verb-agreement prediction task (in which model saws the part of a sentence up to the word to predict), (2) fine-tuning the models on grammaticality judgments of full sentences, and (3) using the models out-of-the box word prediction objective, seeing if the model gives the higher probability to the correctly inflected word than the incorrect one Linzen et al. (2016).

Gulordava et al. (2018), building on Linzen et al. evaluation paradigm, also used naturally occurring sentences on a subject-verb-agreement task, but added also nonce sentence that are grammatical but semantically meaningless, to tease out potential confounds due to semantic cues.

Marvin and Linzen (2018) used instead artificially constructed sentences, noting the limitations of the "naturalistic approach" of using naturally occurring sentences on targeted syntactic phenomena: they are sparse and hard to collect, and they are hard to control for confounds. The constructed sentences approach allowed them to examine a much larger range of specific grammatical phenomena than was possible before. Furthermore, their sentences were

automatically generated using templates, which resulted in a much larger test set, while still balancing the coverage for each phenomenon.

3.1.2 Minimal pairs vs acceptability judgments

Marvin and Linzen (2018) also introduced the **minimal pairs** evaluation paradigm: instead of trying to extract from the model an absolute grammaticality judgment for each sentence in isolation, the model is evaluated on pair of full sentences that differ for only one word, and the model is expected to give an higher probability to the grammatical sentence. This is one possible solution to the problem of how to capture grammaticality from the probability or likelihood score given by language models: the LM probability of a sentence can be a proxy for its acceptability only when considered relatively to another sentence that differs minimally from it, controlling confounding factors like sentence length and lexical content that impact sentence probability (Warstadt et al., 2020).

The minimal pairs evaluation paradigm take advantage of the fact that, while a model's outputted probability of a sentence cannot be taken has an absolute value comparable with any other sentence for grammaticality judgments, the probability scores of minimally different sentences are directly comparable for this purpose (Lau et al., 2016; Warstadt et al., 2019).

On the other hand, the approach of minimal pairs might be more challenging to apply to more complex syntactic phenomena, and approaches that try to extract absolute acceptability judgments from a model have the advantages of being directly comparable with with both native speaker judgments and predictions made in the linguistic literature (Warstadt et al., 2019). Targeted minimal pairs datasets might also be more labor-intensive to produce, when manually collected (Warstadt et al., 2020). Another fundamental difference is that minimal pairs scoring is usually done with the pretrained models out of the box, while acceptability judgments datasets require the models to be fine-tuned on an additional dataset on an acceptability detection task.

Warstadt et al. (2019) introduced the Corpus of Linguistic Acceptability (CoLA), a collection of about 10k of sentences labeled with acceptability judgments taken from the linguistic literature. It was developed to be used with the absolute acceptability judgments approach, and required the models to be fine-tuned on an sentence acceptability task with a subset of the data. This poses the problem of all probing tasks: adapting language models with fine-tuning to perform downstream tasks doesn't necessary reflect knowledge that is already present in the LMs (Warstadt et al., 2020). The models could for instance be overfitting to superficial cues in the fine-tuning training set, rather than actually leverage the linguistic features learned during pretraining. CoLA was also preprocessed to edit out less common words, as detected by checking their occurrence in the British National Corpus (BNC).

A successor of CoLA was the BLiMP dataset (The Benchmark of Linguistic Minimal Pairs), which was developed for the minimal pairs paradigm instead. (Warstadt et al., 2020). The sentences in BLiMP are automatically generated from templates, producing 1000 sentences for each of the covered phenomena, controlling for confounding factors such as sentence length and lexical content. The lexical items are sampled from a 3000 word lexicon, which includes a complex annotation for selectional restrictions to be observed when combining words in a sentence.

As noted in the original BLiMP paper, the minimal pairs paradigm might not be suitable for testing some more complex phenomena, like negative polarity items (NPIs) or island effects: "our implementation of these phenomena is often narrower than the linguistic definition because of the particular constraints described above." (Warstadt et al., 2020) For these phenomena, a more suitable approach might be the factorial design seen in psycholinguistic studies (Sprouse et al., 2016), which also has been later applied for the assessment of LMs (Wilcox et al., 2018; Hu et al., 2020), and that we adopt for the present thesis.

3.1.3 Factorial approaches

A factorial experimental setup can be considered a generalization the minimal pairs approach. A factorial design allows to quantify and tease out the effect of different factors that contribute to a phenomenon, in order to better control confounds (Sprouse et al., 2016, 2012).

The results of a factorial experimental setup can then be analyzed in multiple ways. It is possible to derive measures of statistical significance using a mixed-effects linear regression model, as it's standard practice in psycholinguistics (Sprouse et al., 2016; Wilcox et al., 2018).; or it is possible to obtain an overall accuracy score, by setting up a multi-fold SUCCESS CRITERION, considering an item as accurately scored if multiple inequalities have been satisfied, as done by Hu et al. (2020). This latter case is more suitable for integration into minimal pairs benchmarks like BLiMP, that use accuracy scores, and it is the one we adopt in this thesis.

3.1.4 Sentence probabilities from unidirectional and bidirectional language models

Conventional language models, like those based on long-short term memory units (LSTM), and GPT-2 (Radford et al., n.d.), are unidirectional, they predict the probability of a token using only the previous tokens preceeding it. This allows to estimate the log probability for a sentence W via the chain rule, by summing the log probabilities of each token w_t in the sentence conditioned by the previous tokens (Salazar et al., 2020; Lau et al., 2020; Bengio et al.,

2003):

$$\log P_{uniLM}(W) = \sum_{t=1}^{|W|} \log P_{uniLM}(w_t|W_{<t})$$

On the other hand, BERT (Devlin et al., 2018) is a bidirectional language model, that it's trained to predict the probability of a token given both the tokens preceding and following it in a sequence. The ability to see both left and right contexts is one of the strengths of BERT over models GPT-2, however, this makes it harder to obtain sentence probabilities from its output. An alternative approach that has been proposed, is to estimate a sentence probability by masking one token at a time and summing up the resulting log probabilities of each one, conditioned on the other tokens (Salazar et al., 2020; Lau et al., 2020). The resulting measure is not a true probability (the probabilities of all the possible sentences don't sum up to 1), but a *pseudo-log-likelihood* score (PLL):

$$PLL_{biLM}(W) = \sum_{t=1}^{|W|} \log P_{uniLM}(w_t|W_{\setminus t})$$

This is equivalent to summing up the *surprisal* of each individual token in the sentence. Surprisal is the negative log probability of a word given its context, and intuitively is a measure of how surprising that word is in that context. Surprisal is a well-established measure to predict human incremental processing difficulty, measuring for instance phenomena like garden-path disambiguation effects (Hu et al., 2020).

$$S(w|C) = -\log_2 P(w|C)$$

Where $S(w|C)$ is the surprisal of the word w given its context C.

Salazar et al. (2020) compared the numerical properties of the sentence log probabilities obtained from GPT-2, and the PLL scores obtainaed from BERT, and found that PLL, and found that PLL's summands (the token surprisals) are more uniform across an utterance's length, while GPT-2 displays a left-to-right bias, with high surprisal values at the beginning of the sentence, which decrease as the model sees the rest of the sentence tokens. This is due to the fact that, in a unidirectional language models, at the beginning of a sentence only a few tokens have been seen, therefore it's very hard to predict the following tokens, which therefore tend to have high surprisals. Therefore plots showing the surprisals of each token from a BERT model tend to be "flat" (with all values oscillating approximately within the same range), while GPT-2 models display a descending curve.

This descending curve in the per token surprisals in unigram LMs is intrinsic to the fact of being able to see only the left context of a token, and tends to hide the spikes in surprisals of individual tokens due for instance in

ungrammaticality. For this reason, the per token surprisals of bidirectional models like BERT, help to differentiate fluency from likeliness (Salazar et al., 2020).

Lau et al. (2020) proposed to normalize by sentence length the sentence scores obtained from GPT-2 (LP) and BERT (PLL), by dividing them by a penalty term which depends on the sentence length measured in number of tokens:

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

Where LP is the log probability of the sentence s. Lau et al. (2020) found better overall across models results with $\alpha = 0.8$. However, Salazar et al. (2020) found that dividing by this penalty term is beneficial only for unidirectional models like GPT-2, because it has a dampening effect on the descending curve of the token surprisals discussed above. For bidirectional models like BERT, instead this normalization is detrimental, and the unnormalized PLLs tend to give better estimates of sentence probabilities. With these measures, Salazar et al. (2020) found that BERT outperforms GPT-2 on the BLiMP benchmark. We found confirmation of this fact in the results of preliminary experiments for this thesis, and for this reason in our comparisons we'll score GPT-2 with PenLP and BERT-based models with the PLL to estimate sentence acceptability.

3.2 Results from previous related work

Multiple works have reported that transformer-based language models are able to learn most syntactic and semantic phenomena with under 100M tokens of pretraining data, however these models still lag behind human performance on some more complex syntactic and semantic phenomena, including island effects (Zhang et al., 2021; Rogers et al., 2020).

Chowdhury and Zamparelli (2018) used yet another methodology for estimating grammaticality judgments, in this case not on transformer-based models but on Recurrent Neural Netwoks (RNNs). Instead of measuring the probability of the whole sentence, they measured the probability of the question mark punctuation, as a proxy for the RNNs gap expectation in filler-gap dependencies. They found that RNNs were not able to learn some types of island constraints. However Wilcox et al. (2018) notes that their results might be due to their experimental setup rather than actual limitations of the models representations.

Warstadt et al. (2020) found that GPT-2 showed near-chance performance on island effects on the BLiMP benchmark. However, Salazar et al. (2020) found that, on the same BLiMP test set, RoBERTa large (an optimized vari-

ant of BERT) learns island effects close to a human level, with accuracy scores of 83.4% from RoBERTa (whereas humans score at 84.9%). They suggested that suggesting that the discrepancy with previous results with GPT-2 was due to the different numerical properties of the output of unidirectional and bidirectional language models.

However, as it has been pointed out in the original BLiMP paper (Warstadt et al., 2020), high performance on test items fro island effets in the BLiMP benchmark might not be wholly indicative of the models real abilities, since with the minimal pair paradigm it's possible to capture a very narrow definition of island effects. Knowledge of island effects phenomena is better and more comprehensively evaluable with a factorial design like (Wilcox et al., 2018).

Wilcox et al. (2018) tested Recurrent Neural Networks (RNNs), and Long Short-Term Memory RNNs (LSTM) in particular, but not transformer-based LMs, on island effects, with a factorial experimental paradigm. They found evidence that these models learned wh-islands , adjunct islands, and complex NP islands, but not subject islands. However, their study was primarily focused on Filler-Gap dependencies, and their experimental definition of island effects is based on a reduction of the in filler-gap effects, specifically the fact that, while in the absence of an island structure a model expects to find a gap if there is a filler present, this expectation is significantly reduced when there is also an island structure. This factorial experimental definition of island effects is different from the one used in psycholinguistic study on human subjects (Sprouse et al., 2016; Bondevik et al., 2021), in which it is measured as the residual acceptability difference between between the baseline condition (absence of island structure, and absence of a filler-gap dependency), and island violation, after the contributing effects of presence of island structure filler-gap dependency have been subtracted away.

Hu et al. (2020) tested LSTMs and GPT-2 (but not bidirectional LMs) on several syntactic phenomena, including long-distance dependencies, using a 2 x 2 factorial experimental setup. They measured the probabilities of critical sentence regions instead of full-sentence probabilities. They defined factorial success criteria for each phenomenon, enabling them to obtain accuracy percentage scores as end results. This makes their approach suitable withing benchmarks like BLiMP, that also produce a final accuracy score. They found that GTP-2 scored almost 80% accuracy on long-distance dependencies, but they did not compared it with a human reference score.

Chapter 4

Experimental setup

4.1 Test suite design

In building test suite for our experiments, we largely followed the factorial design of the target material in Italian that was administered by Sprouse et al. (2016) to human subjects (that we discussed in section 2.4), but we increased item number of 50 from the original 8 (by comparison, previous work on language models like Wilcox et al. (2018) used about 20 items per suite).

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, one per CONDITION, as in item 4.1.2, covering the combinations of two FACTORS: presence or absence of an island structure (the STRUCTURE factor), and and presence of a short or long-distance dependency (GAP-POSITION factor). Therefore, each sentence in an item varies minimally from the other 3 to exemplify one these four CONDITIONS: SHORT-NONISLAND, Long-NonIsland, Short-Island, and Long-Island. The first three are supposed to be acceptable sentences, while the last one is supposed to be unacceptable.

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 4.1.2 has four sentences given by the combination of these two factors.

In item 4.1.2, 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.)

Example 4.1.1. 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 _]?

We have two types of adjuncts islands. In temporal adjuncts, the island structure is a temporal adjunct clause headed by temporal adverbs like "*dopo, prima, quando, mentre*" ("*aftern, before, when, while*"). Conditional adjuncts are headed by "*se*" ("*if*").

Example 4.1.2. ADJUNCT ISLANDS (TEMPORAL)

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 printed the illustration after the author sent to the publisher?')

Example 4.1.3. ADJUNCT ISLANDS (CONDITIONAL)

a. SHORT-NONISLAND:
Chi _ dice che il pilota aumenterà la velocità dell'aereo?
('Who says that the pilot will increase the plane speed?')

b. LONG-NONISLAND:
Che cosa dici che il pilota aumenterà _?
('What do you say that the pilot will increase?')

c. SHORT-ISLAND:
Chi _ infrangerà il muro del suono [se aumenterà la velocità dell'aereo]?
('Who will break the sound barrier if he increases the plane speed?')

d. LONG-ISLAND:
*Che cosa il pilota infrangerà il muro del suono [se aumenterà _]?
('What will the pilot break the sound barrier if he increases?')

Example 4.1.4. 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?')

We followed the experimental setup of (Sprouse et al., 2016), in order to compare our results with theirs. Therefore the factorial definition of subject islands is modified to the other three island types. The GAP-POSITION factor for the other three island types varies between MATRIX (short-distance dependency) and EMBEDDED[1] (long-distance dependency), while for subject islands it varies between EMBEDDED-OBJECT (there is an extraction of the object, or from the object, of the subordinate clause) and EMBEDDED-SUBJECT (there is an extraction of the subject or from the subject of the subordinate clause). As discussed in (Sprouse et al., 2016) this causes the non-parallel lines

[1]The "matrix" clause is the main clause, while the "embedded" clause is a subordinate clause.

in the plots to have slopes in opposite directions, because the EMBEDDED-OBJECT-NONISLAND sentences have the long-distance dependency (therefore lower acceptability), while the Embedded-subject-NonIsland sentences have short-distance dependencies.

The rationale for this change is to minimize potential textual ambiguity in signaling the gap locations of the oblique (PP) arguments, due the fact that Italian does not allow *preposition stranding*. Preposition stranding is the possibility of having a preposition *stranded* (separated from) its object. For example, in the English sentence *I know **who** you bought the painting **from***, "from" and "who" are stranded. Instead, Italian forces "*pied-piping*" (another metaphorical technical term introduced by Ross (1967), like that of "islands", as mentioned in chapter 2), the fact that the prepositional phrase must be continuous (the wh-word *brings along* the preposition): *So **da chi** hai comprato il quadro* ("*I know from who you bought the painting*").

Example 4.1.5. SUBJECT ISLANDS

a. EMBEDDED-OBJECT-NONISLAND:
Chi pensi che la decisione avvantaggi _?
('Who do you think the decision benefits?')

b. EMBEDDED-SUBJECT-NONISLAND:
Chi pensi che _ abbia assunto la decisione?
('Who do you think took the decision?')

c. EMBEDDED-OBJECT-ISLAND:
Di chi pensi che [la decisione del sindaco] avvantaggi gli interessi _?
('Who do you think the mayor's decision benefits the interests of?')

d. EMBEDDED-SUBJECT-ISLAND:
*Di chi pensi che [la decisione _] avvantaggi gli interessi degli agricoltori?
('Who do you think the decision of benefits the farmers' interests?')

4.2 Sentence acceptability estimates

We assessed uni-directional and bi-directional transformer-based Italian language models. For the uni-directional models (based on GPT-2), we scored sentences using the log probability given as output by the model, which is calculated by multiplying the estimated probabilities of each token t_i of a sentence s using the previously seen left tokens $t_{<i}$ as context(Lau et al., 2020; Bengio et al., 2003):

$$\log P(s) = \sum_{i=0}^{|s|} \log P_{GPT2}(t_i|t_{<i})$$

For bi-directional models, based on BERT, we estimated sentence probabilities using a pseudo-log-likelihood score (PLL), obtained by summing the conditional log probabilities of each token, using as context the rest of the tokens in the sentence (both left and right), as in Salazar et al. (2020) and Lau et al. (2020):

$$PLL(s) = \sum_{i=0}^{|s|} \log P_{BERT}(t_i|t_{<i}, t_{>i})$$

We used these two scores as a relative estimate of sentence acceptability, for the purpose of comparing similar sentences (varying across controlled conditions) within an item like (15).

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, normalized to the same scale all the scores of each model across the four wh-dependency island effects phenomena of a particular test suite.

4.3 Tested models

All the four Italian models tested have the same numer of parameters (about 110M) They include one unidirectional language model, GePpeTto (De Mattei et al., 2020), based on the GPT-2 architecture (Radford et al., 2018); and three bidirectional language models based on the BERT architecture (Devlin et al., 2018).

GePpeTto[2] is a cased model trained on a 13.8GB Italian corpus, which correspond to about **2.2B tokens**, from Wikipedia and ItWac corpus. The model's size corresponds to GPT-2 small, with 12 layers, a hidden layer size of 768 units, and 12 attention heads, for a total of 117M parameters. The vocabulary size is 30k.

BERT[3]): a BERT-base cased model trained on a 13GB Italian corpus, corresponding to **2B tokens**, from a Wikipedia dump, and the OPUS corpus. The model s composed of 12-layer, with a hidden layer size of 768 units, 12-heads, and a total of 109M parameters.

BERT XXL [4]: it has the same architecture of the above BERT model, but it was trained on a 81GB corpus, corresponding to **13B tokens**, with included the OSCAR corpus, in addition to the same Wikipedia dump and OPUS corpus used for the above BERT model.

[2]https://huggingface.co/LorenzoDeMattei/GePpeTto
[3]https://huggingface.co/dbmdz/bert-base-italian-cased
[4]https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

GilBERTo [5]: an uncased model trained on 71GB of Italian text (**11.2B** tokens) from the OSCAR corpus. It uses the RoBERTa base architecture (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer and Stoyanov, 2019) with a model size of: 12-layer, 768-hidden, 12-heads, a vocabulary of 32k sub-word units, and a total of about 110M parameters.

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

Chapter 5

Results

5.1 Accuracy on factorial scores

In Table 5.1 we report the results of the evaluation of the four Italian models described in section 4.3, on our test suites covering four island effect types. The accuracy scores are based on the success criterion for each item, defined as $DD > 0$, where the DD score is calculated as described in section 2.4 with the formula 2.4. To recap, the DD score is obtained by subtracting from the total effect (the difference in acceptability between the baseline sentence and the ungrammatical one) the effects of dependency length (gap-position) and structure (presence or absence of an island construct), to obtain a measure quantifying only the violation of the island constraint (that is, the island effect).

We see that the BERT-based models outperform GPT-2 in the overall scores, and this confirms findings in previous work like Salazar et al. (2020). GPT-2 is outperformed by BERT even when trained on a similar amount of data (around 2B tokens), with a difference of over 5 points in accuracy (from 81.5% to 86.5%), which increases (although with diminishing returns) in models trained on larger data. BERT XXL, trained on 13B tokens, achieves 88.5% accuracy, while GilBERTo, trained on 11B tokens, performs the best at 91%.

Looking at performance in individual island types, we can see that whether islands seems to be the easiest one to learn, with all models achieving 100% accuracy except BERT at 96%. This suggests that the mastering of this type of island (in the configurations we tested) reaches 100% accuracy somewhere after 2B and 11B tokens of training data.

Instead, subject islands, at least in the forms captured in our test suite, seem to be the hardest type of island to learn, with still room for improvement after the 11B of training data of GilBERTo, that reaches 94% accuracy. This type of island also shows a direct correlation with the accuracy scores and the amount of training data: the models trained on 2B of tokens reached

Island type	**Factorial score**	**GePpeTto** LP	**BERT** PLL	**BERT XXL** PLL	**GilBERTo** PLL
Whether	$DD > 0$	**100**	96	**100**	**100**
Adjunct	$DD > 0$	**98**	70	64	72
ComplexNP	$DD > 0$	46	96	**100**	98
Subject	$DD > 0$	82	84	90	**94**
Average		81.5%	86.5%	88.5%	**91%**

Table 5.1: Accuracy percentages for GPT-2 and BERT Italian models, on a test suite of 50 items per phenomenon developed for the present thesis. For the GPT-2 model we estimated the sentence acceptability with the sentence log probability (LP), while for the BERT-based models we used the sentence pseudo-log-likelihood (PLL). The DD score is calculated by subtracting the structure and long-distance effects from total effect, as in (Sprouse et al., 2016). The success criterion for each factorial item of 2 x 2 sentences is $DD > 0$.

and accuracy of 82% and 84%, while the models trained on 11-13B tokens achieved an accuracy of 90% and 94%. This seems to indicate that at least some type of island effects are among the hardest syntactic phenomena to be grasped by current transformer-based language models, at least when trained on ungrounded raw text in a self-supervised task without explicit linguistic supervision. All of these four models are trained a minimum of 2B tokens, which is already an order of magnitude more than the amount of data (100M tokens) at which they learn most (but not all) syntactic and semantic phenomena (Zhang et al., 2021). It is also possible that the current targeted syntactic benchmarks like BLiMP are not challenging enough, and overestimate these models actual linguistic abilities, as has been argued by Newman et al. (2021).

Another possible explanation for the fact that these models seem to need massive amounts of data to learn subject islands, much more than for other syntactic phenomena, is that this correlates with Italian speakers finding subject islands weaker than other types. Indeed, Sprouse et al. (2016) found that subject islands with wh-dependencies are the type of island that scored the lowest (indicating "weaker" island effects) on human Italian subjects (Figure 5.1), and that subject islands with rc-dependencies even showed almost no island effects on human subjects.

For adjunct and complex NP islands, the results show a mixed picture. For adjuncts islands, performance seems to be correlated not with training set size (at least beyond 2B tokens), but with architecture, with GPT-2 outperforming the other models and reaching 98% despite being trained on 2.2B tokens.

Complex NP islands, on the other hand, show the opposite picture: it seems to be the GPT-2 architecture to be struggling with them (with a very low 46% score, which could be considered below chance), while the other models

Model	**Adjunct**	**Complex NP**	**Wh**
GPT-2 (Warstadt et al 2020)	91	72	77
GPT-2-large	90.2	72	79.1
GPT-2-medium	**91.6**	72.3	77.9
GPT-2	91.3	68.8	82.2
BERT-base-cased	88.1	56	66.2
BERT-large-cased	86.3	67.4	69.4
RoBERTa-base	84.9	75.3	76.2
RoBERTa-large	86.9	**82.3**	**88.2**

The score on the first row are taken from the original paper Warstadt et al. (2020). The model parameters sizes are the following: GPT2 124M, GPT-2-medium 355M, GPT-2-large 774M all trained on about 8B tokens from the WebText corpus. BERT 110M, BERT-Large 340M, both trained on 3.2B tokens. RoBERTa-large 125M, RoBERTa -large 355M, both trained on 32B tokens.

Table 5.2: Accuracy results on some extraction islands phenomena in the BLiMP English test suite for BERT and GPT-2 models.

score at 96% or higher.

5.1.1 Cross-linguistic comparison with BLiMP scores

For cross-linguistic comparison, we reproduced the results on the English BLiMP benchmark reported by Warstadt et al. (2020); Salazar et al. (2020) for the three island types it has in common with our test suites: wh-islands, adjuncts, and complex NP islands and we report our results in Table 5.2.

Preliminarily, we note that the format of the BLiMP tests and ours is quite different, so this comparison can be only indicative and tentative. Regarding the sentence score normalization, we note that on the BLiMP test suites, it makes no difference whether to use the acceptability score normalization with a penalty on sentence length (which results in the PenLP score defined by Lau et al. (2020)), because each minimal pair is controlled to have sentences of exactly the same length. However, the BLiMP minimal pairs paradigm allows only for a narrower definition of island effects, that does not really capture adequately the phenomenon.

We observe that, as in the results for Italian models, the GPT-2 architecture seems to be favored in adjunct islands, in which it outperforms by at least 5 percentage points the BERT-based models on the BLiMP test suite. For the other two types of islands (complex NP and Wh) we see that the results don't seem to be directly comparable to ours, and this is probably due to the very different experimental setup and difference definition of these two types of islands captured by BLiMP and by our test suites.

5.1.2 Using factorial sentences as minimal pairs

We also explored the possibility of deriving minimal pairs from each factorial item, by building three pairs from the unacceptable sentence and the three unacceptable ones. However this is not feasible, because in this way the resulting pairs are "less minimal", since confounds are not controlled for lexical and syntactic variations This could lead even to cases where the baseline least complex sentence (SHORT-NONISLAND condition) would receive a very low likelihood score for being very short and having less semantic cues, compared to longer sentences, therefore presenting spikes of surprisals for some tokens, that are less predictable given the less informative context. In a factorial setup, on the other hand, confounds, even some implicit unexpected ones, tend to be subtracted out in a properly factorial score (although also in this case is important to have minimally varying sentences within each item, which is tricky task for some type of sentences).

We report in Table 5.3 the accuracy scores obtained in such a way, deriving minimal pairs from the factorial items. Each acceptable sentence of an item like item 4.1.2 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 whether island type, the acceptable Short-NonIsland sentence gets scored higher than the unacceptable sentence (Long-Island) 92% of the times by the GPT-2 model with the LP sentence acceptability estimate, and 96% of the time when the sentence score is normalized by length as in Lau et al. (2020). The GilBERTo model (trained on 11B tokens) reaches the top overall average accuracy, with a score of 91.7% which is similar to the one obtained with the factorial scoring.

5.2 Qualitative analysis

In this section we compare our results obtained from transformer-based language models with those from human subjects collected by Sprouse et al. (2016). We complement this with an analysis of the per-token surprisals given by the models for some of the items in our test suites.

With this qualitative analysis we try to explain the results in Table 5.1. Specifically, the fact that (1) for adjunct islands, there is a 6+ points drop in accuracy by BERT XXL (the model trained on the largest amount of data), compared to BERT and GilBERTo. That (2) for subject islands, there is a progressive increase in accuracy, which more or less correlates with amount of training data, from BERT (2B tokens), to BERT XXL (13B tokens), to GilBERTo (11B tokens).

The fact that (3) complex NP islands, there is a slight increase in perfor-

Phenomenon	Sentence form	Gpt2 (it) LP	Gpt2 (it) Pen LP	BERT (it) LP	BERT (it) Pen LP	GilBERTo (it) LP	GilBERTo (it) Pen LP
Wh-whether	Short-N.I.	92	96	**100**	**100**	**100**	**100**
	Long-N.I.	**100**	**100**	**100**	**100**	**100**	**100**
	Short-I.S.	64	98	90	96	**100**	**100**
Wh-adjunct	Short-N.I.	**96**	92	94	90	90	82
	Long-N.I.	**98**	86	68	42	76	54
	Short-I.S.	96	98	**100**	98	96	94
Wh-complex np	Short-N.I.	90	92	**100**	**100**	**100**	**100**
	Long-N.I.	**100**	42	96	92	**100**	98
	Short-I.S.	38	88	**100**	**100**	**100**	**100**
Wh-subject	Short-N.I.	**100**	94	28	8	68	36
	Long-N.I.	**100**	98	86	56	98	86
	Short-I.S.	40	56	62	60	72	**76**
Average		84.5%	86.7%	85.3%	78.5%	**91.7%**	85.5%

Table 5.3: Accuracy results for Gpt-2 and BERT Italian models when using the sentences in a factorial design as minimal pairs. Evaluated on a test suite of 50 items per phenomenon developed for the present thesis.

mance of BERT-based models (from 96% to 98% to 100%) correlated with the amount of training data of each model. The fact that (4) GPT-2 shows a complete drop in performance (to 46%) for this type of islands. The fact that (5) for whether islands, BERT (trained on 2B tokens) is the only model to score below 100% accuracy, at 96%, and therefore performance for this type of island seems to reach 100% somewhere between 2B and 11B tokens.

In Fig. 5.1 we see the original plots from Sprouse et al. (2016), on results collected from Italian subjects for wh-dependencies islands. Figures 5.2 to 5.3[1] show the plots we obtained for the same types of island constructs and wh-dependencies, with our test suite (that has a format similar to that of Sprouse et al.), on three Italian language models: the GPT-2-based GePpeTto (De Mattei et al., 2020)[2], BERT (base-xxl version)[3] and the RoBERTa-based GilBERTo[4]. For comparison, we also show in Fig.A.9-A.12 plots on the scores given by the three above language models on the original test suites by Sprouse et al. (2016).

The y axis represents the structure condition (island or non-island). The x axis (for whether, adjunct, and complex NP islands) represents the dependency distance: short-distance extraction from the matrix clause, or long-distance extraction from the embedded clause. The x axis for subject islands,

[1]In these plots the scoring measures are indicated with the terminology by Lau et al. (2020): LP stands for log probability, which for BERT models it's actually the pseudo-log-likelihood (PLL), while the sentence score normalized with a penalty term based on sentence length is indicated with PenLP.

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

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

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

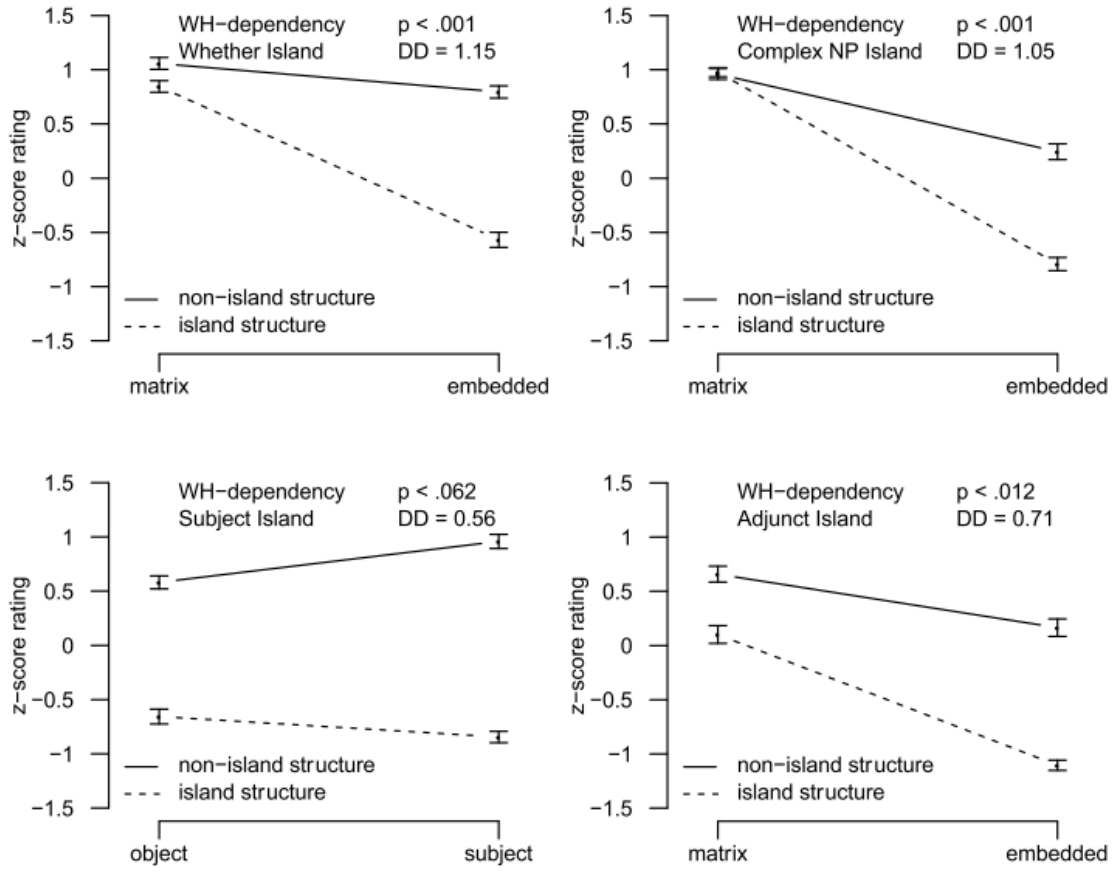


Figure 5.1: Plots of average acceptability scores from humans taken from (Sprouse et al., 2016)

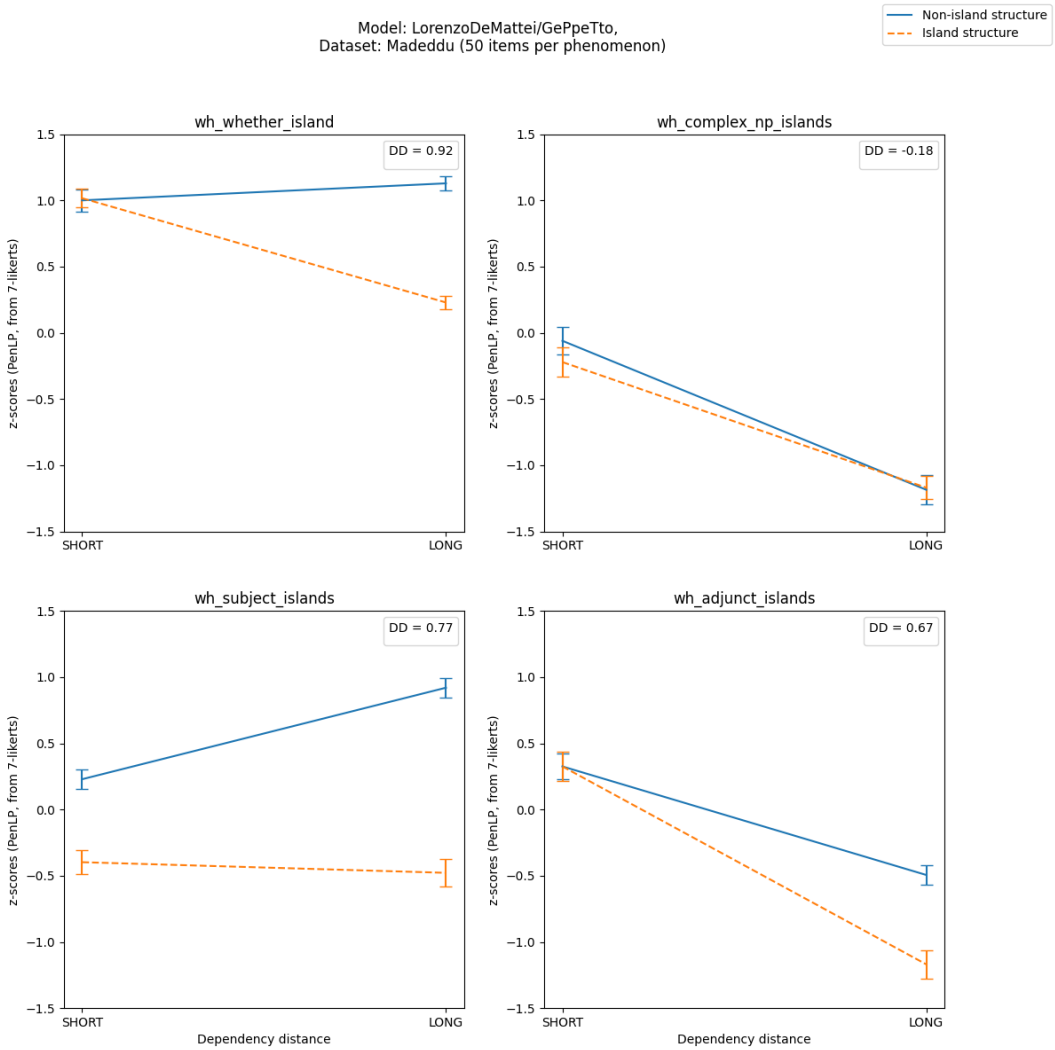


Figure 5.2: Plots of average acceptability scores from GePpeTto, on the test suites developed for the present thesis.

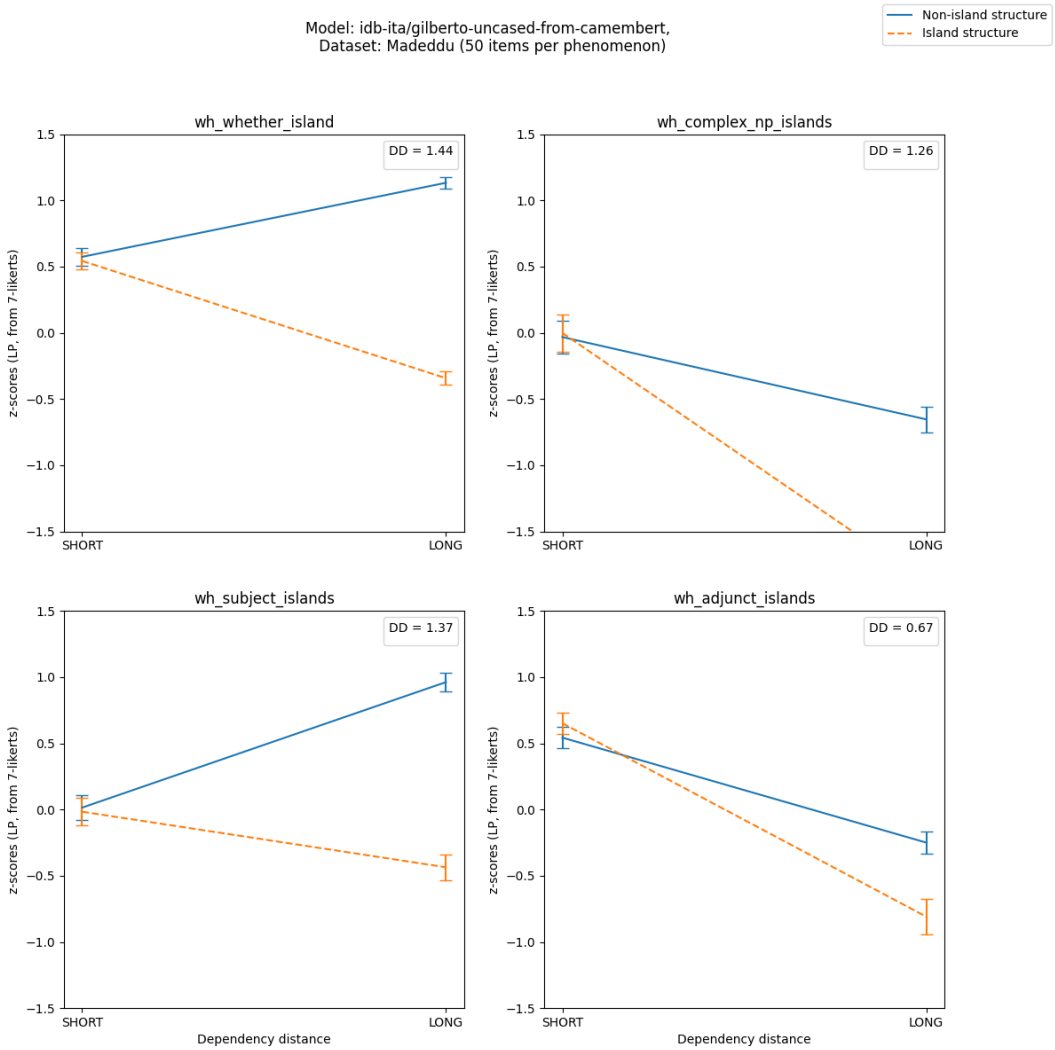


Figure 5.3: Plots of average acceptability scores from GilBERTo, on the test suites developed for the present thesis.

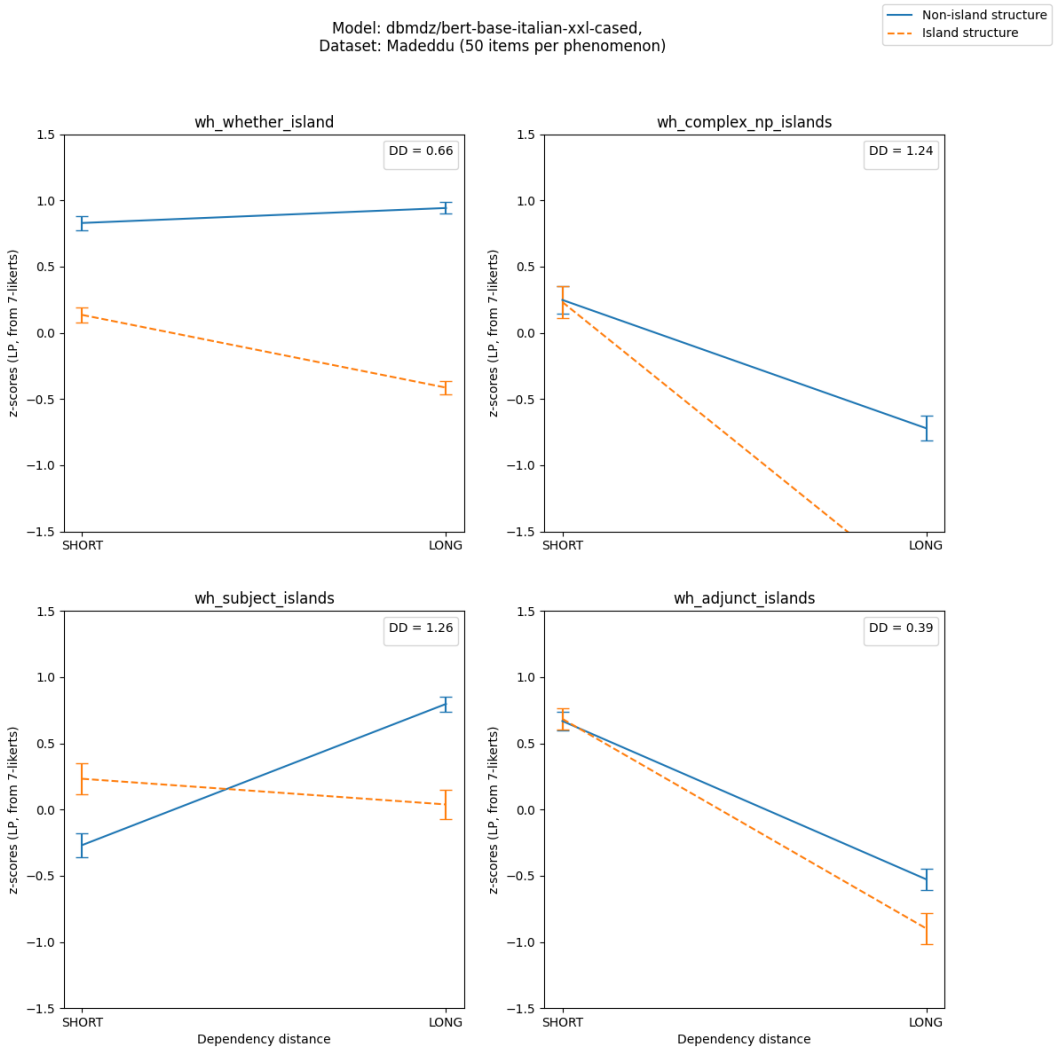


Figure 5.4: Plots of average acceptability scores from BERT XXL (13B of training tokens), on the test suites developed for the present thesis.

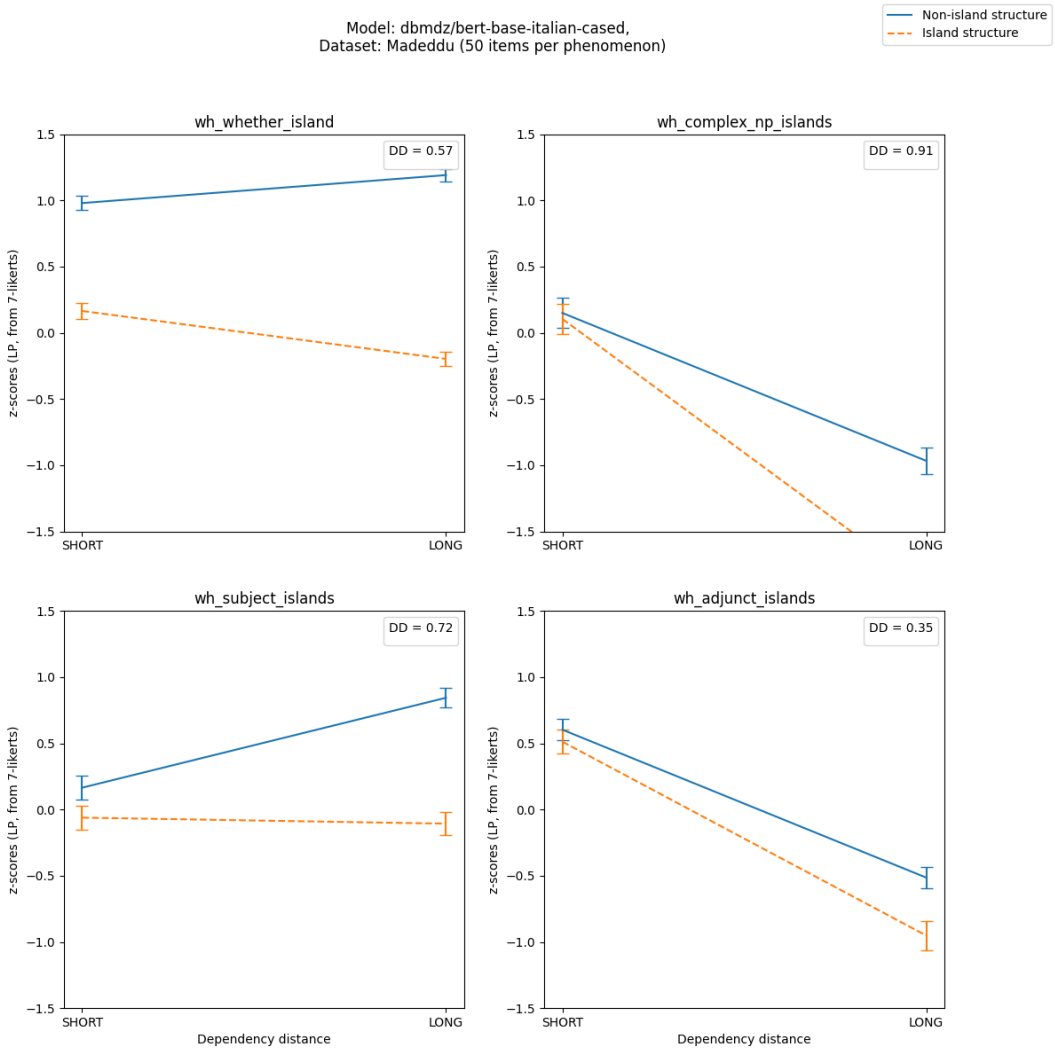


Figure 5.5: Plots of average acceptability scores from BERT (2B of training tokens), on the test suites developed for the present thesis.

instead, as discussed in section 4.1, represents extraction of the object from the embedded clause, or extraction of the subject from the embedded clause. The "non-island" line, marked in solid blue, connects the values for the two sentences without the island structure. The orange dotted line is the "island line", connecting the values for the two sentences with the island structure.

We can see that the plots for whether (top-left plot of each figure), adjunct (bottom-right), and subject islands (bottom-left) from the GPT-2 model (Figure 5.2) show some similarities with those obtained on human subjects (Figure 5.1), but also some significant differences.

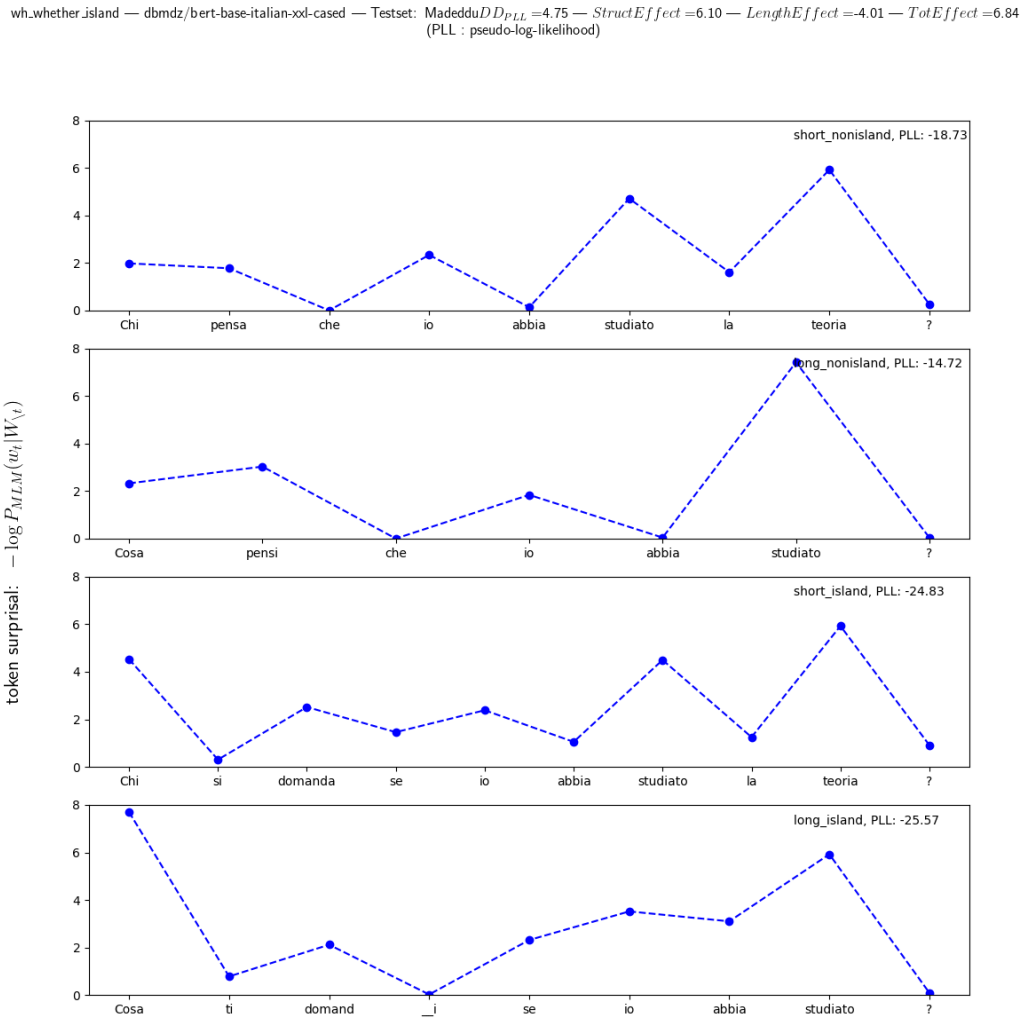
5.2.1 Whether islands

For whether islands, one difference between the plots from human scores and those from the models (Figure 5.2-Figure 5.5), is that for whether islands the line connecting the SHORT-NONISLAND and the LONG-NONISLAND goes upwards instead of downwards like for humans; that is, on average the language models give the LONG-NONISLAND a higher acceptability estimation score (the pseudo-log-likelihood, PLL), than the SHORT-NONISLAND (and this trend remains also when the models score the original test suite administerd by Sprouse to humans, diplayed in Figure A.9-Figure A.12).

To explain why, we can have a look at Figure 5.6, which shows the surprisals of each token in one of the whether islands items. Plots showin the surprisal for each token in a sentence are a kind of "zoom" into the sentence acceptability estimates, since, in the case of BERT-based models, the sentence score (the pseudo-log-probability) is just the (negative) sum of the token surprisals.

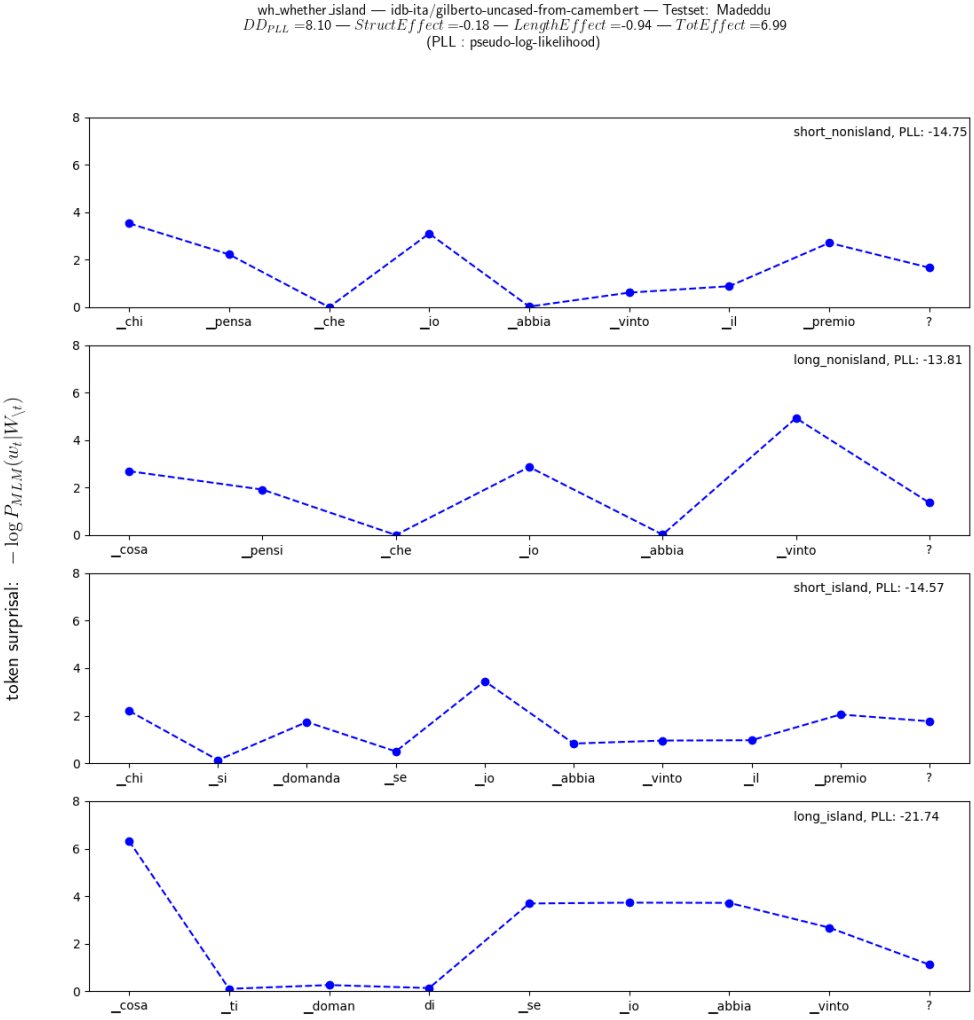
In Figure 5.6, we see that the SHORT-NONISLAND sentence ("*Chi pensa che io abbia studiato la teoria?*") has the additional tokens "*la teoria*", which significantly adds to the total sum of surprisal. The same phenomenon can be see in Figure 5.7, where the phrase "*vinto il premio*" (found in the SHORT-NONISLAND) sums up to slightly more surprisal than the word "*vinto*" alone (found in the LONG-NONISLAND). However, in general we can't just assume that a longer sentence will get a lower acceptability scores, because having more content words can give more semantic cues, making it easier to predict each of them when it's masked during the scoring, lowering the surprisals (as can be seen in Figure 5.12, discussed in the next section).

Back to the question of why BERT is the only model not to reach 100% accuracy on whether islands, there seems to be no clear explanation for this fact looking at the token surprisals of the same item across all the models, as shown in Figure 5.9-5.11. One thing we notice is that the baseline acceptable sentence (SHORT-NONISLAND) ("*Chi pensa che io abbia aumentato l'affitto?*") ("*Who thinks that I have raised the rent?*") gets a lower acceptability in BERT



Token surprisal plots like this are a kind of "zoom" into the sentence acceptability estimates, since, in the case of BERT-based models, the sentence score is just the (negative) sum of the token surprisals. At the top of the figure there is the factorial DD score on the whole item (one that is > 0 the item is considered scored accurately). There are also the factorial measures of length and structure effect, and the total effect. Note that these are raw values obtained from the (negative) sum of the token surprisals, while the values seen in plots like in Figure 5.3 are normalized across all the sentence scores for all the phenomena given by a particular model.

Figure 5.6: Token surprisal for the four sentences of one of the whether island items, as scored by the Italian BERT XXL.



Note that the tokens are lowercased because GilBERTo is available only in the uncased version, otherwise uppercase words would be split, increase the surprisals, and introduce confounds.

Figure 5.7: Token surprisal for the four sentences of one of the whether island items, as scored by GilBERTo.

XXL (Figure 5.10) than BERT (Figure 5.9). The trend of the surprisal plots seems to be almost the same between BERT and BERT XXL, with a few difference. We can notice the strange phenomenon that BERT XXL gets a high surprisal (much higher than BERT) on the words "*affitto*" ("*rent*") and "*io*" ("*I*"). Again this seems another case of semantic phenomena having a prevailing confounding effect, that could mask the effect of syntactic ones.

5.2.2 Adjunct islands

GilBERTo has a much lower accuracy score (72%) than GPT-2 (98%) for adjunct islands (Table 5.1), and this is the only island type in which it doesn't outperforms or at least come very close to both of the other two models. From the plots it seems that this is due to the LONG-NONISLAND sentences getting on average too low of an acceptability, so the expected LENGTH-EFFECT that will be subtracted from the total effect (formula 2.4) is really large (Figure 5.3).

Looking at the token surprisals for one of the adjunct island items (with a *conditional* adjunct island construct introduced by "*se*", "*if*"), we can see in Figure 5.12 that the LONG-NONISLAND (second sentence from the top) gets really high spikes for the phrase "*il sindacato indirà*" ("*the union will call*"), while in the case of the SHORT-NONISLAND (first sentence from the top), the phrase "*il sindacato indirà lo sciopero*" ("*the union will call the strike*") gets a much smaller surprisal on the word *sindacato* ("*union*"), seemingly because the presence addition of the word for *strike* (which, mutually, also gets almost no surprisal for the same reason) makes it much more likely to predict when masked. However, these spikes are also present (albeit a bit ameliorated) in the LONG-ISLAND sentence (last at the bottom), so this effect would be subtracted out by the factorial scoring. The thing that keeps the DD score negative (and therefore renders this item inaccuratly scored) is that the LONG-NONISLAND sentence has also a surprisal spikes for the phrase "*dici che*" ("*you say that*"), and this seem to be due to syntactic reasons: the model seems to consider less likely this kind of long-distance dependency construct ("*che cosa dici che il sindacato indirà?*") than another long-distance dependency construct where there is a island violation ("*che cosa gli edicolanti chiuderanno i battenti se il sindacator indirà?*"). We can conclude then that the worse accuracy score that GilBERTo gets in adjuncts islands (Table 5.1) seems to be indeed due to an unrobust learning of this type of adjunct island constraints by GilBERTo.

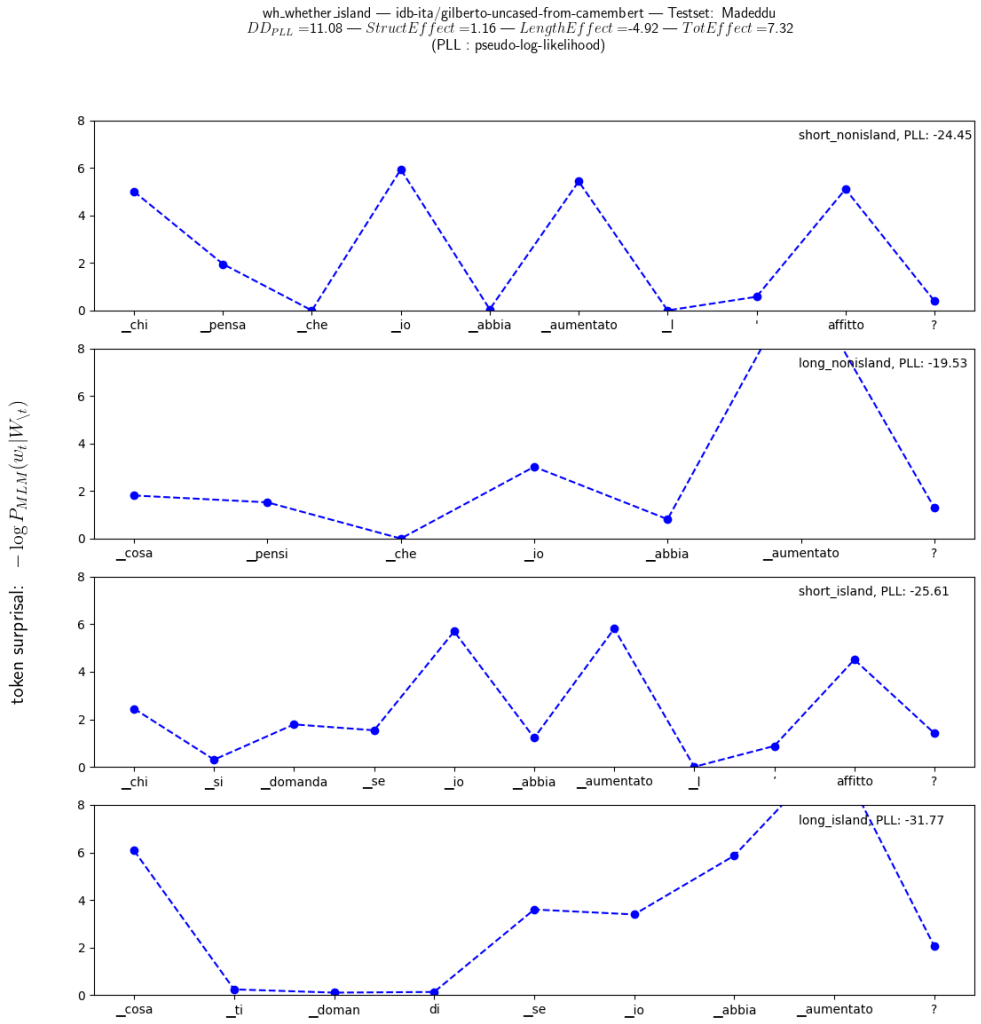


Figure 5.8: Token surprisal of a whether island item, as scored by GilBERTo.

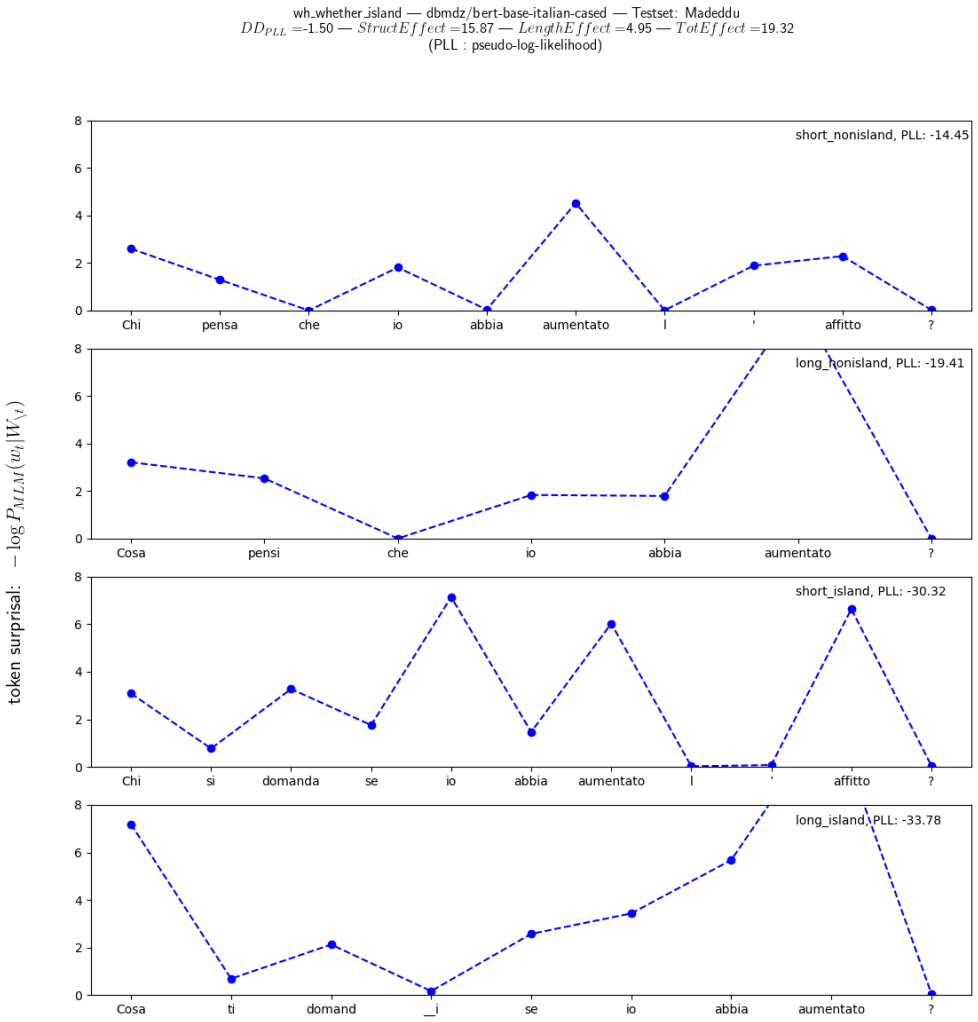


Figure 5.9: Token surprisal of a whether island item, as scored by the Italian BERT.

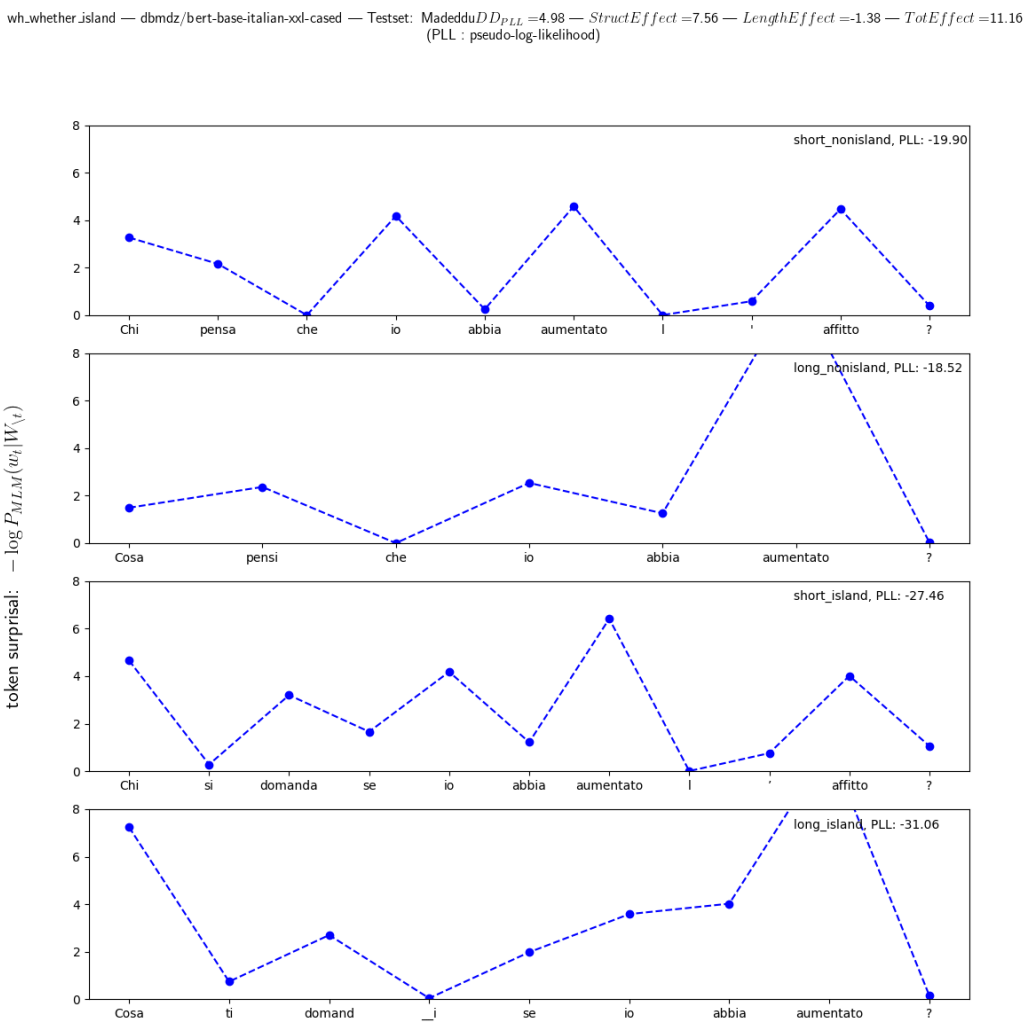


Figure 5.10: Token surprisal of a whether island item, as scored by the Italian BERT XXL.

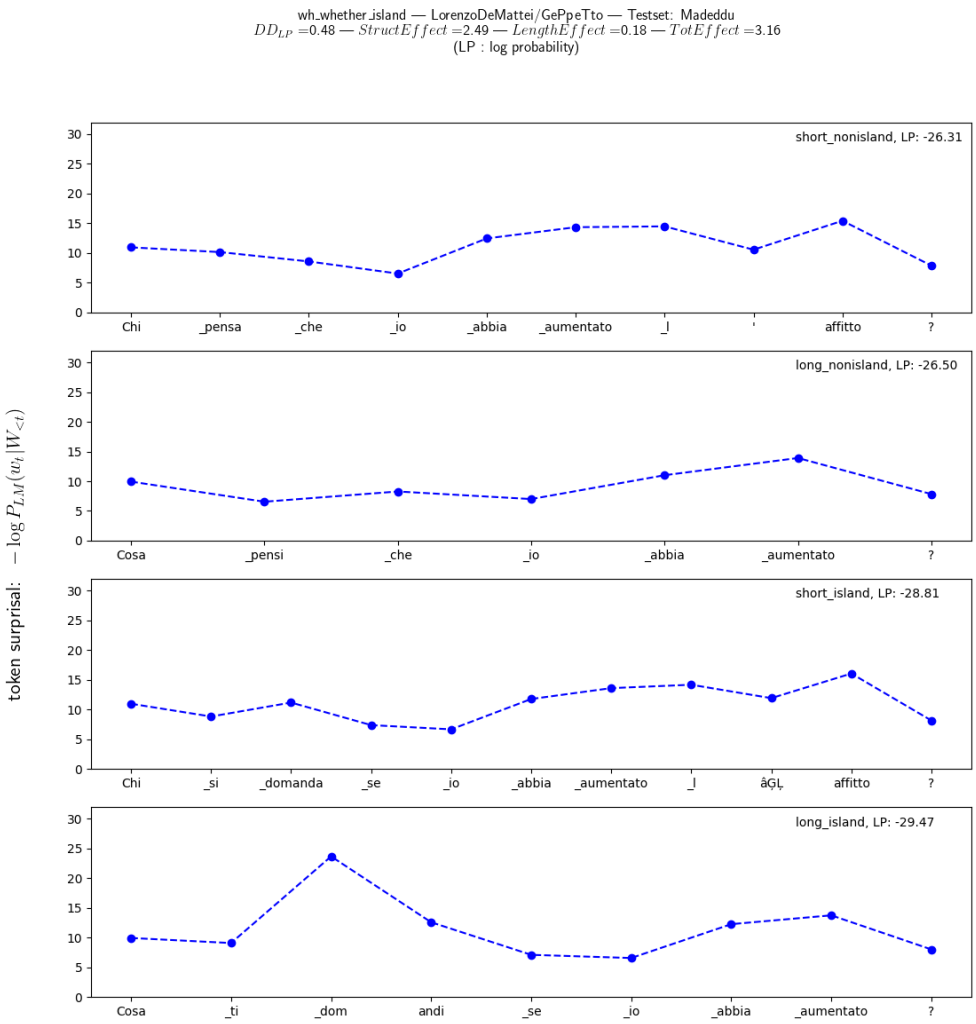


Figure 5.11: Token surprisal of a whether island item, as scored by GePpeTto.

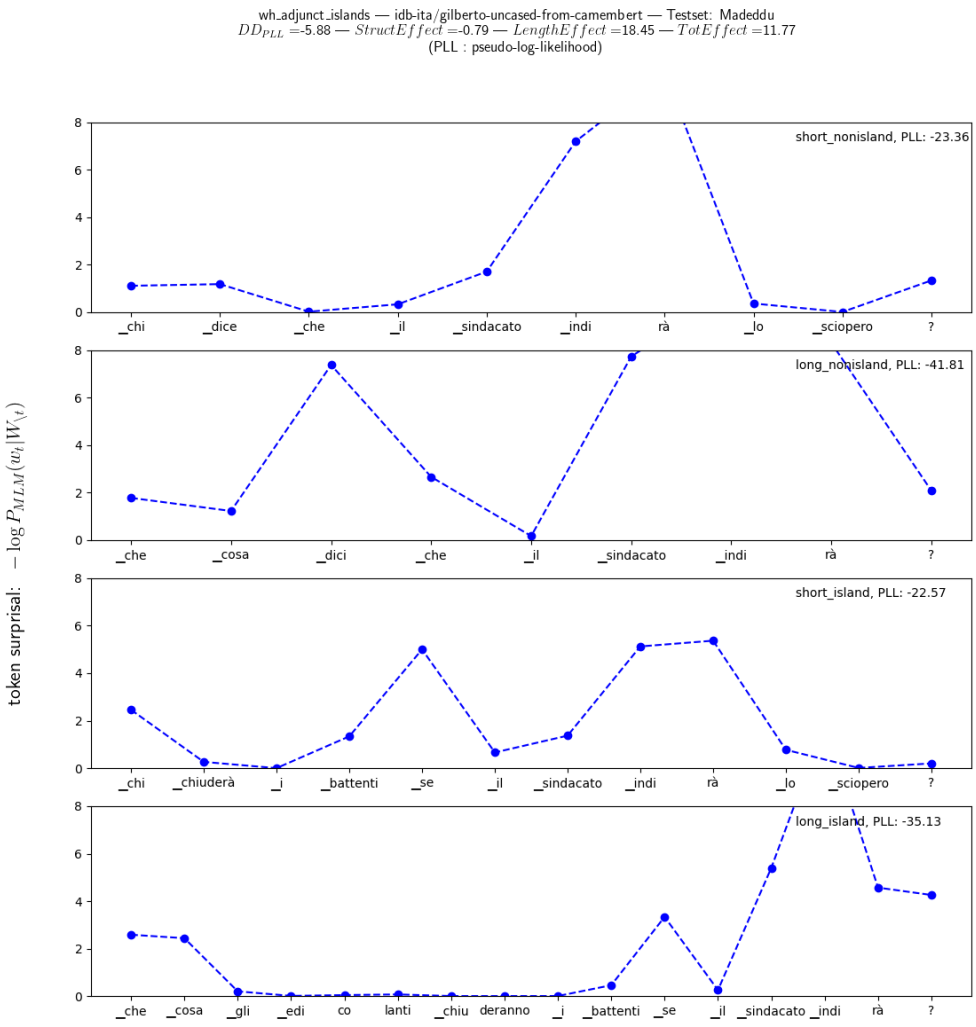


Figure 5.12: Token surprisal for the four sentences of one of the adjunct island items, as scored by GilBERTo.

Also in Figure 5.12 (an adjunct island item scored by GilBERTo), we can notice what seems to be causing the SHORT-ISLAND sentences to have a higher acceptability estimate than SHORT-NONISLAND ones, and therefore the island and non island lines to "cross" in Figure 5.3 (plot at the bottom-right). The SHORT-NONISLAND too has the phrase "*il sindacato indirà lo sciopero*", but it has also additional content words that give semantic cues and therefore further lower the surprisals, with the phrase "*chiuderà i battenti*".

GPT-2 seem to have learned adjunct island constraints more robustly than the two BERT-based models, as mentioned above. Looking at Figure 5.2 it seems this is due to the higher gap in acceptability given by GPT-2 than BERT and GilBERTo (Figure 5.3). In the case of GPT-2, the gap it's almost 1 standard deviation (these are normalized z-scores), while it's about half that

for the BERT-based modes (Figure 5.3).

We can complement our analysis so far, with a comparison of the token surprisals for the same sentences across the four models we tested. In Figure 5.13, 5.14, 5.15, 5.16 we can see token surprisals plots from BERT, BERT XXL, GilBERTo and GePpeTto on an adjunct island item.

We observe that a prevailing factor influencing the models probabilities, and therefore the surprisals and the final sentence acceptability estimates, seems to be how much a model has learned the probabilities of certain collocations. In particular we can see a difference between BERT and BERT XXL, seemingly due to the different amount of data they have been trained with (2B vs 13B tokens).

In Figure 5.13, in the top sentence, we can see that BERT (the model trained with the least data), has not learned a strong association between "*emendamento*" ("*emendament*") and "*la camera boccerà*" ("*the lower chamber will reject*"), and for this reason the word gets a surprisal spike that highly influences the final sentence score. However, in the third sentence, we can see that the same word ("*emendamento*") does not get the same spike, likely for the presence, far back at the beginning of the sentence, of the word "*legge*" (*law*), for which probably the model already has learned a strong association with "*emendamento*".

Looking in Figure 5.14 at the plots for the same sentences from BERT XXL (trained with billions more data), we can see that the spikes in the first sentence for *"camera"* ("*lower chamber*") and "*emendamento*" ("*emendament*"), have been much lowered compared to the previous model, likely because after 13B tokens of training data BERT has finally learned the association between these two words.

Factors like this can tilt the final factorial scores of an item, affecting its accuracy, and playing as a confound that masks these models responses to the syntactic phenomena we are interested. For these reasons, the sentences in a factorial item should be strictly balanced lexically, ideally having the same words (just in different order, as it is often the case in the BLiMP test suites), or at least have the same semantically full content words, with variations only in functional words. However, this is a non-trivial task, and designing targeted linguistic benchmarks in such a way, and for a comprehensive number of linguistic phenomena it's challenging and potentially very expensive. The approach adopted in works like BLiMP, to automatically generate the items from templates, is not necessarily less demanding, since for more complex phenomena there is also an increase in the the complexity of the templates and of the annotation schema for the selectional preferences of each word in the sampling lexicon such as the one used in BLiMP.

Back to our initial question, on why BERT XXL, the model trained with

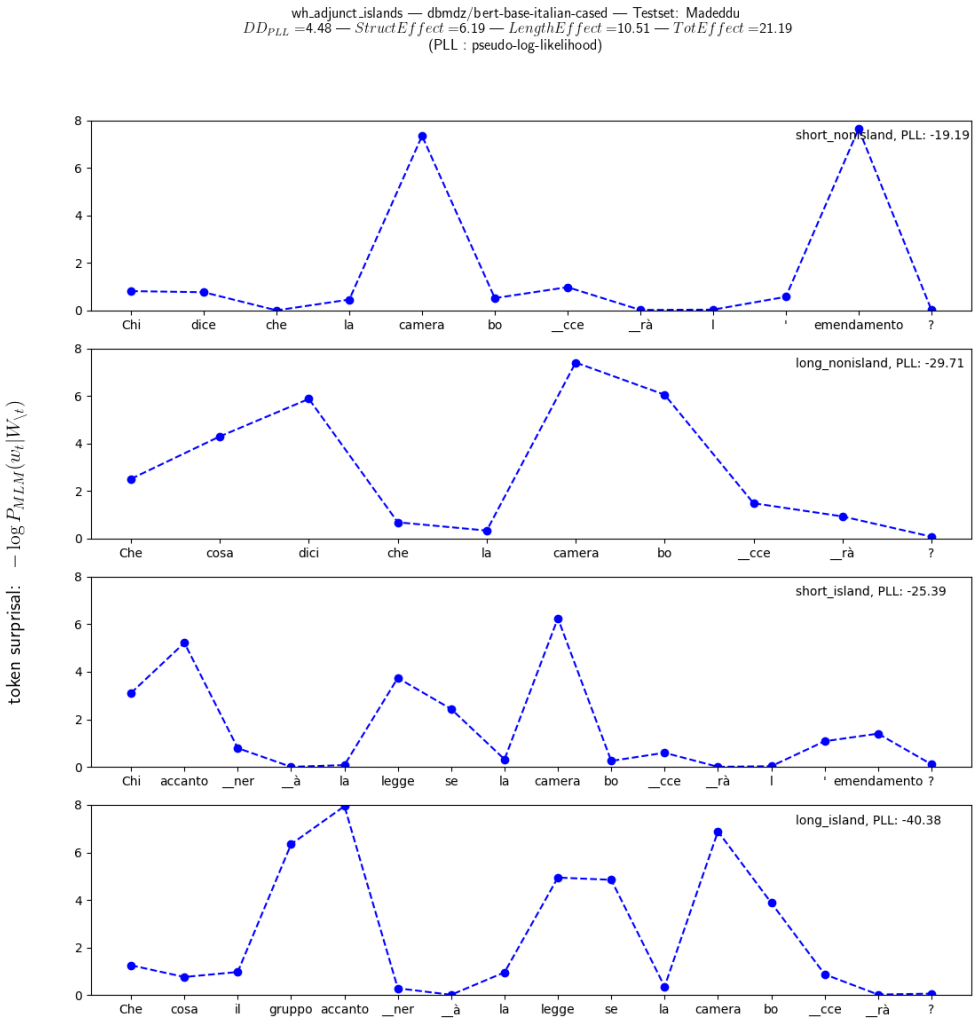


Figure 5.13: Token surprisal of an adjunct island item, as scored by the Italian BERT.

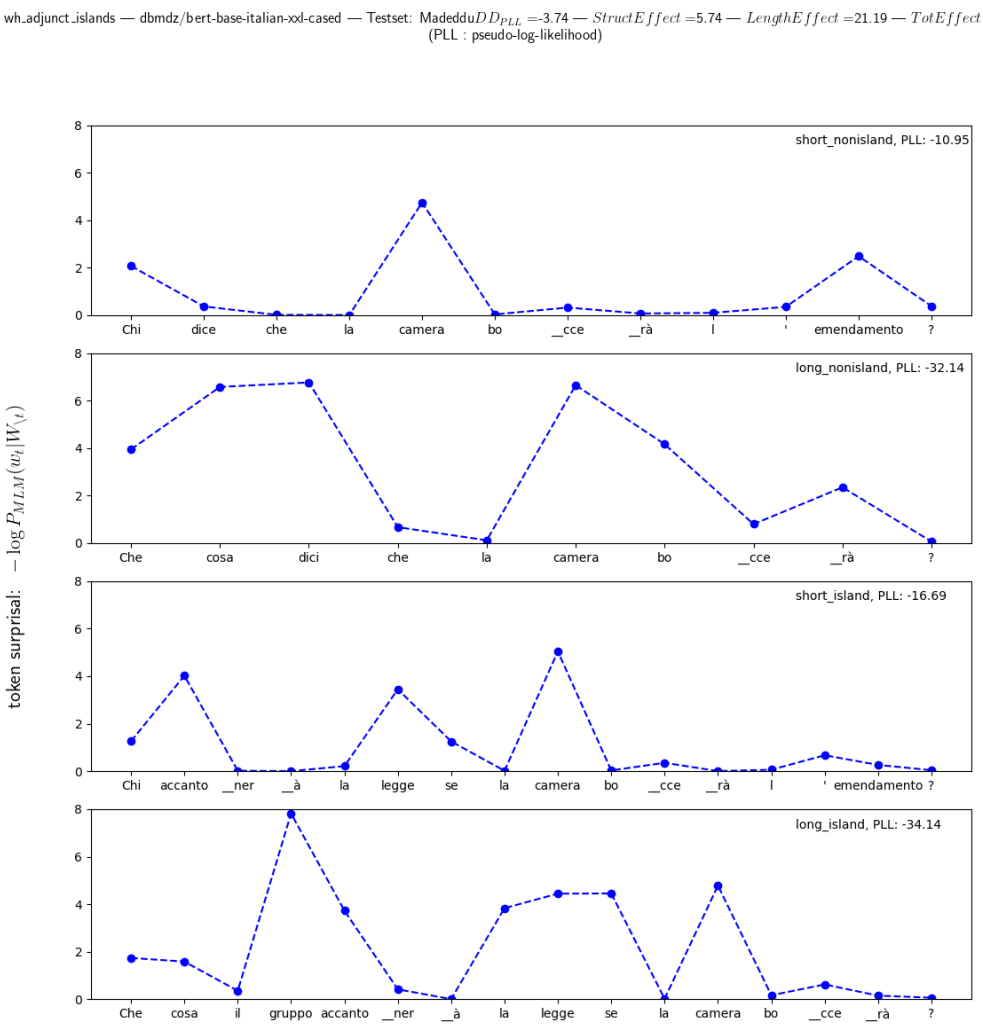


Figure 5.14: Token surprisal of an adjunct island item, as scored by the Italian BERT XXL.

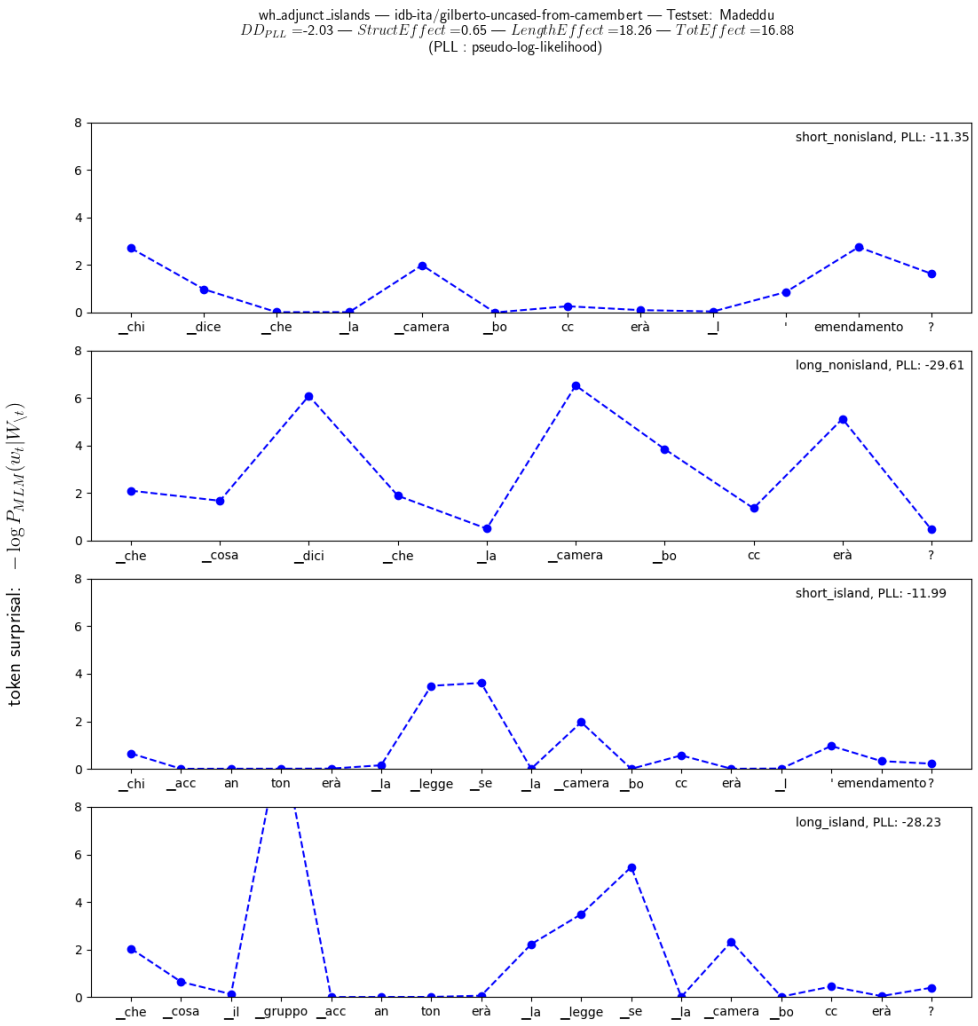


Figure 5.15: Token surprisal of an adjunct island item, as scored by GilBERTo.

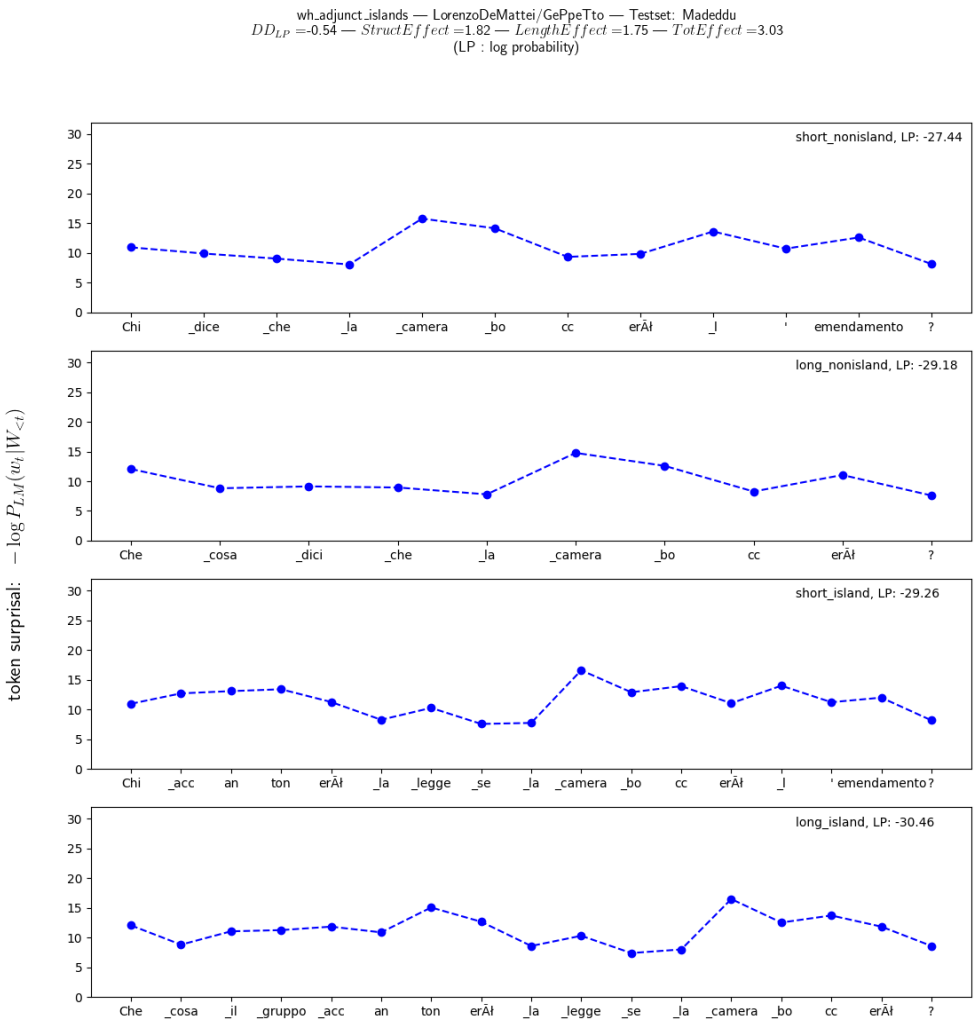


Figure 5.16: Token surprisal of an adjunct island item, as scored by GePpeTto.

the most data (13B tokens), has a 6+ points drop in accuracy for adjunct islands, compared with BERT and GilBERTo, looking at the token surprisals in Figure 5.13, 5.14, and 5.15, we can see that for the bottom ungrammatical sentence ("*Che cosa il gruppo accantonerà se la camera boccerà*") ("*What will the group set aside if the lower chamber rejects?*"), all models have a spike for the word *"se"* ("*if*"), which is a critical region that could indicate that the model detects a grammatical influency. All three models also show a surprisal spike at this word in the 2nd sentence from the bottom ("*Chi accantonerà la legge se la camera boccerà l'emendamento?*") ("*Who will set aside the law if the lower chamber rejects?*"), but these spikes are lower compared to the previous sentence. This could indicate that the models correctly respond with higher surprisals to the presence of an island construct, and also correctly, they further increase this surprisal when there is a violation of the extraction constraint given by this construct, showing that they have learned this type of island effects. However, in these plots we can notice that these surprisal spikes at the word *"se"* ("*if*"), are not clearly reflected in the final PLL scores of each sentence, because of the presence of larger surprisal spikes from other words, likely due to more prominent semantic effects.

Looking at the plots for the average scores in Figure 5.4, we can see that the lines between sentences show very close average acceptability values between sentences with and without and island structure and a dependency of the same distance, with a relatively small slope difference between the two lines (when they are parallel, it indicates that the model is insensitive to that type of island effect). Therefore we can hypothesize that the lexical semantic confounding factors observed in the previous factors can easily shift the overall factorial DD score above or below the defined threshold of the success criterion for accuracy ($DD > 0$). This could partially explain why all the three BERT-based models showed relatively poor performance (64-72%) in this type of islands. GPT-2, on the other hand, which scored at 98%, shows a more marked separation and difference in slope between the two linesFigure 5.2 .

5.2.3 Complex NP islands

Conversely from the adjuncts islands, GPT-2 seems to struggle with complex NP islands (46% accuracy in Table 5.1), compared to BERT (100% accuracy) and GilBERTo (98% accuracy). This is also reflected in the factorial plots showing the average sentence acceptability scores: the one for BERT (Figure 5.4) and GilBERTo (Figure 5.3) have slopes similar to the plots from human scores (Figure 5.1), while GPT-2 stands out as different (Figure 5.2).

We see that, for complex NP items (top-right plot of each figure from Figure 5.1 to Figure 5.5), the long non-island sentences scored by the GPT-2 model with a lower acceptability than by the human subjects, and on the right edge the blue line and orange line converge to the same low value. This does

not happen in the BERT XXL (Figure 5.4) and GilBERTo models (5.3). This might be due to the GPT-2 model being more sensitive to more marginal sentences ending with a gap followed by a question mark, as in the long-distance dependency sentences tested for this type of island ("*Che cosa pensi che Gianni avrebbe comprato _?*", en: "*What do you think Gianni has bought?*").

Back to the question of why BERT XXL achieves 100% accuracy in complex NP islands, going 4 points above BERT, we can have a look at the token surprisals plots in Figure 5.18 and Figure 5.19 for a sentence that BERT XXL scores correctly (DD score = 17.75) and BERT incorrectly (-0.21).

We observe that the difference in the final DD score, seems largely driven by the better acceptability that BERT XXL gives to the LONG-NONISLAND sentence (second from the top in the plots), which results in a much lower expected length effect (7.95 vs 18.86) in the final factorial calculation. Looking at the sentence in question *"Cosa hai rivelato che il consiglio di amministrazione avrebbe redatto?"* (*"What did you reveal that the board of directors had drafted?"*) We see that BERT assigns a larger surprisal value to the region "*Cosa hai rivelato*" ("*What did you reveal*"), and this could in fact be due to a syntactic phenomena, with BERT XLL having learned the less common construct that this phrase introduces here. We also see that BERT gives a much larger surprisal to the word "*amministrazione*" (literally "*administration*", with stands for "*directors*" in the multi-word expression (MWE) equivalent to "*board of directors*"), and this could instead be due to BERT not having learned yet (even after 2B tokens of pretraining data) the collocation "*consiglio di amministrazione*".

5.2.4 Subject islands

Comparing the plots from BERT XXL (Figure 5.4) with those on humans (Figure 5.1), we can see that the in the BERT XXL plot for subject islands, the two lines cross, unlike in those for humans. This is due to the fact that BERT XXL (on average) scores with very low acceptability the baseline sentence (SHORT-NONISLAND condition), in fact lower than the other three sentence conditions. So the variation from it to any of the other three sentences sees and increase in acceptability. But the overall DD score is still positive (hence the 90% accuracy BERT XXL gets on this island type, albeit less then the 94% by GilBERTo), because the increase in acceptability from SHORT-NONISLAND to LONG-ISLAND (a-d in formula 2.4) is still lower than the increase expected from the sum of the two other factors (length and structure), therefore this less-than-expected increase can be attributed to a super-additive island effect.

Back to the question of explaining the correlation between the amount of model training data and the performance of these models on subject islands,

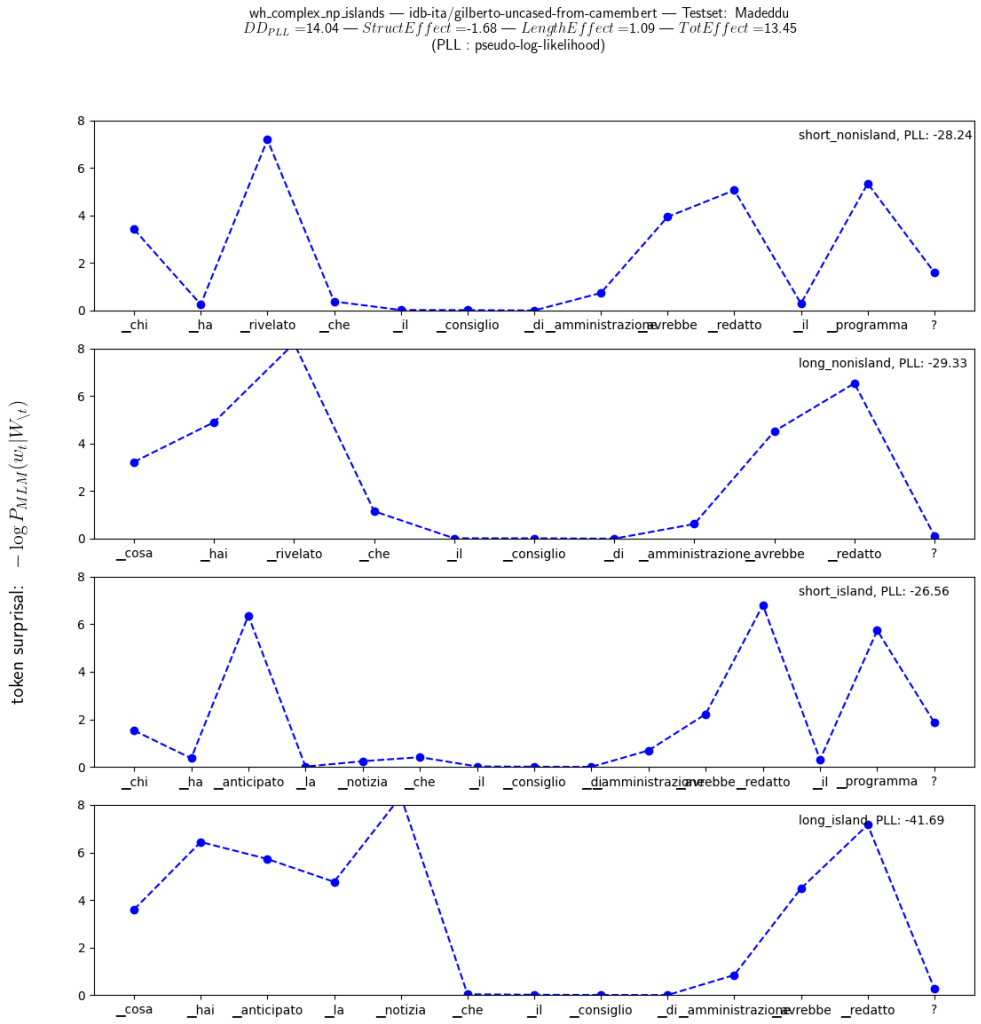


Figure 5.17: Token surprisal of a complex NP island item, as scored by GilBERTo.

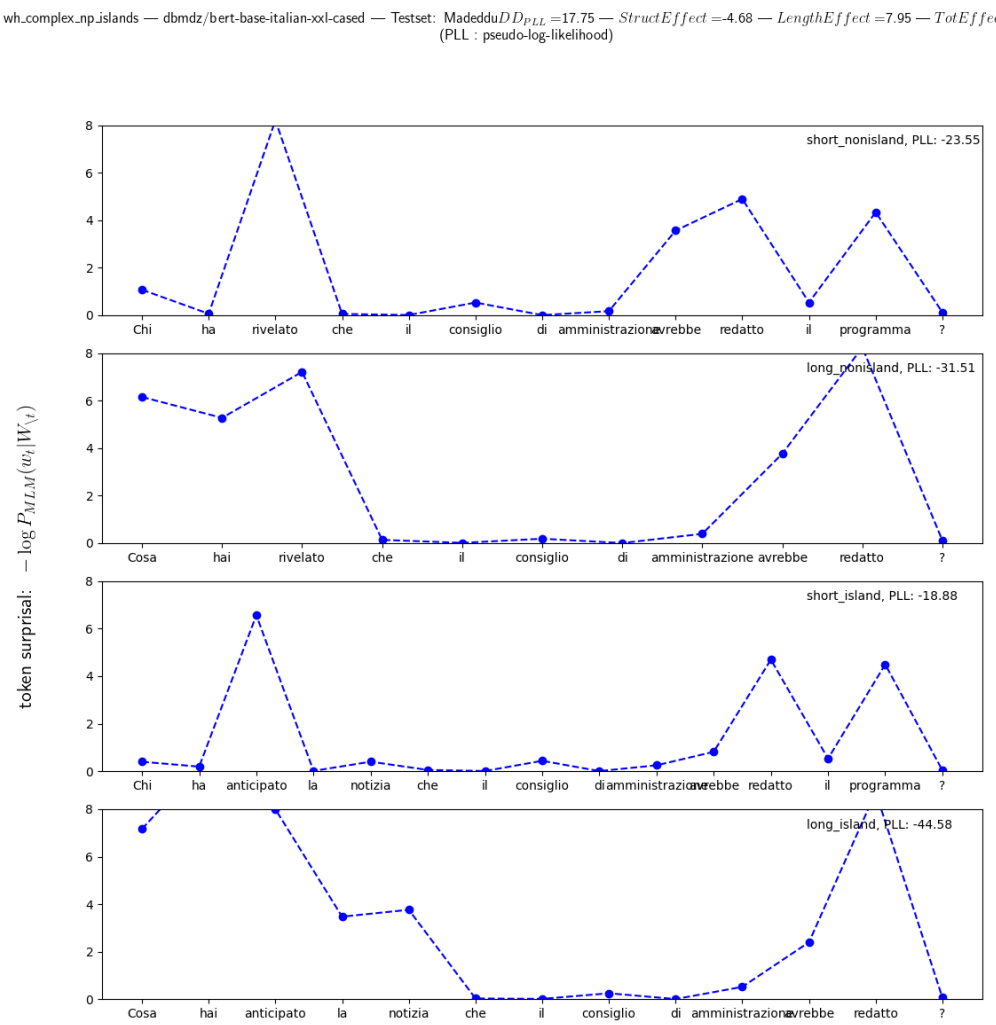


Figure 5.18: Token surprisal of a complex NP island item, as scored by the Italian BERT XXL.

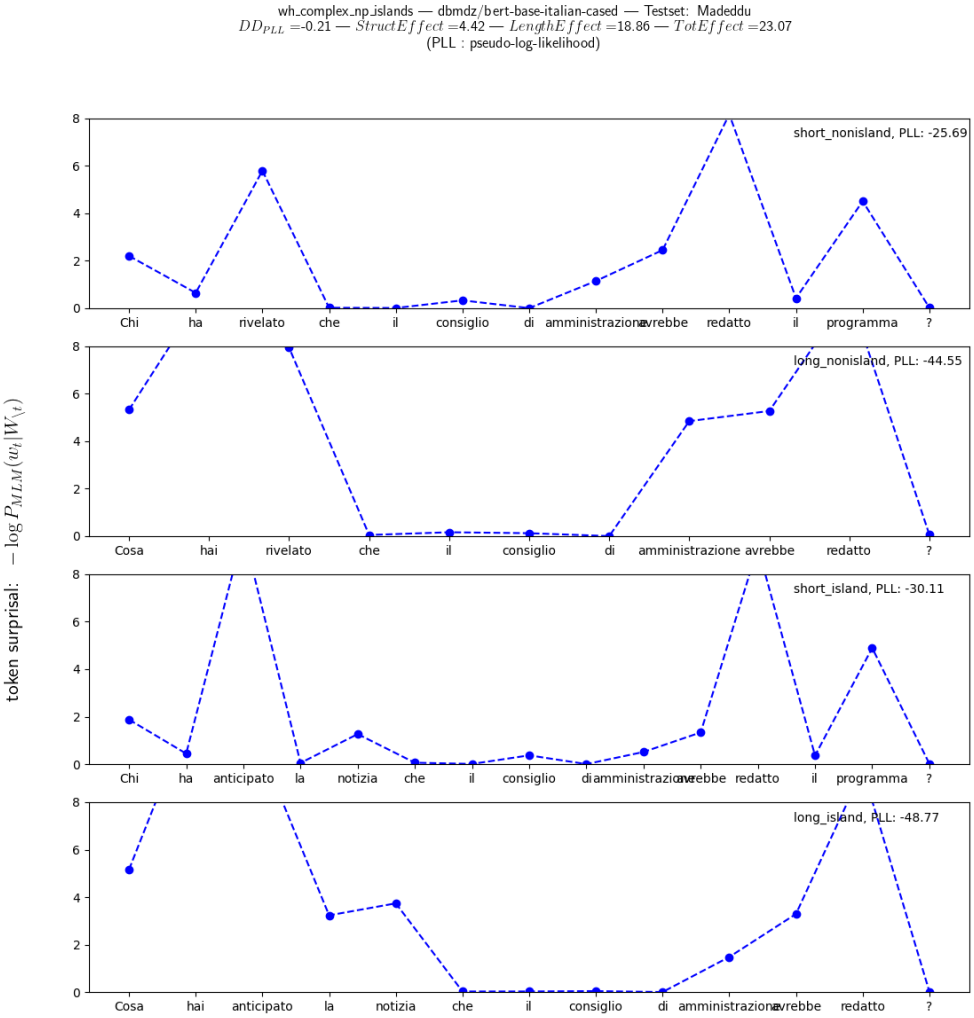


Figure 5.19: Token surprisal of a complex NP island item, as scored by the Italian BERT.

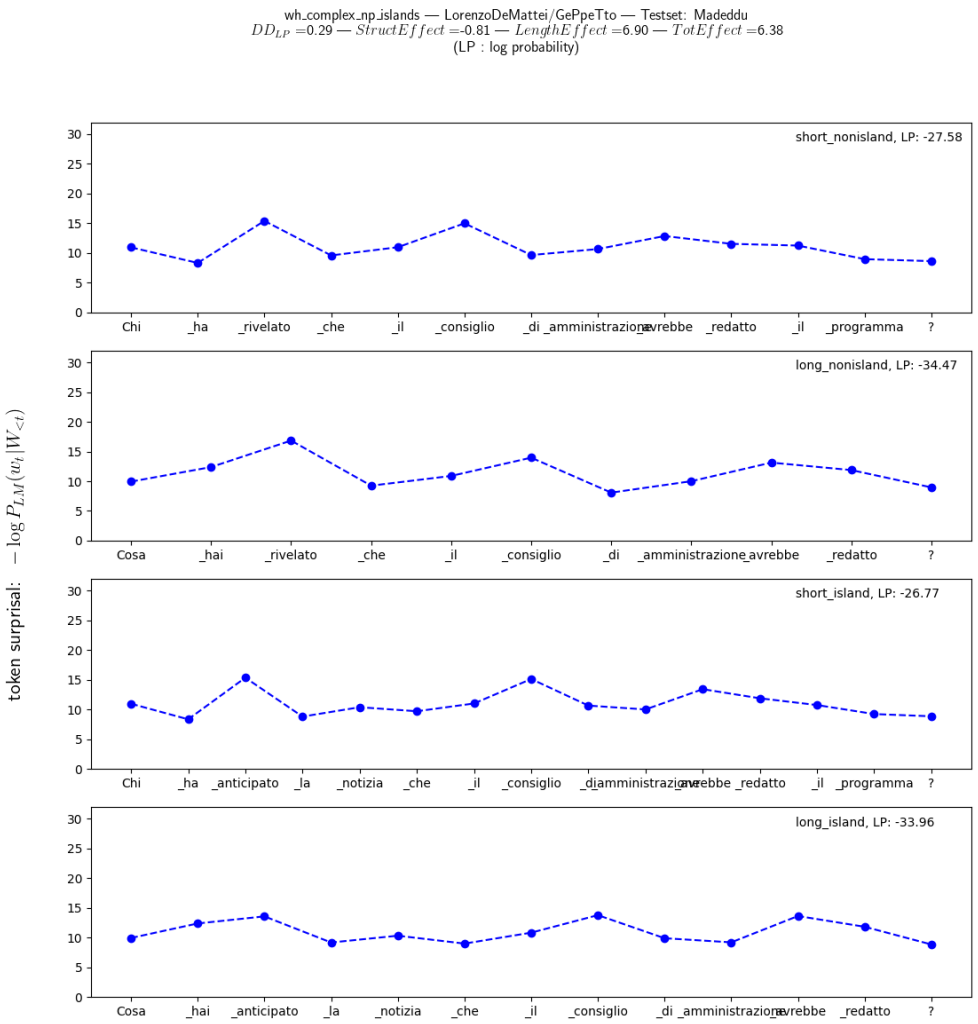


Figure 5.20: Token surprisal of a complex NP island item, as scored by GePpeTto.

we can have a look at the token surprisal plots for the same test item across the four models (Figure 5.21-5.24). In Figure 5.22, we see that BERT XXL gives a better acceptability to the baseline sentence than the BERT model (Figure 5.22), but this seems to be due again to lexical reasons, because BERT XXL gives much smaller surprisals to the words "*foto*" ("*photo*") and to the OOV word "*scandalizzato*" ("*scandalized*"). We also observe that in the two bottom sentences ("*Di chi pensi che la foto dell'artista abbia suscitato l'indignazione?*", "*Di chi pensi che la foto abbia suscitato l'indignazione del pubblico?*") ("*of whom do you think that the artist picture caused the outrage?*", "*of whom do you think that the picture cause the outrage of the audience?*") we notice that BERT, compared to BERT XXL, gives very high surprisal to an apostrophe punctuation marking the determiner elision and followed by a OOV word ("*l'indignazione*") ("*the outrage*").

5.2.5 Other observations

We found confirmation of what found by Salazar et al. (2020), that for bidirectional models like BERT, it's better not to normalize by sentence length the sentence score (the pseudo-log-likelihood), contrary to what proposed by Lau et al. (2020). As observed by Salazar et al. (2020), a sentence final acceptability score, as obtained by the sum of the surprisal in a bidirectional LM, has a "flat" trend, independent to its length, and it is determined instead by surprisal spikes of a few tokens, often for semantic reasons.

We can see the degrading effect in performance, due to normalizing by sentence length, comparing the plots for BERT XXL without normalization (Figure 5.4) and with normalization (Figure A.1 in the Appendix). This is shown as the divergence of the models plots, from those from the human scores, are exacertbated when using sentence length normalization. The two lines (connecting the average values of each sentence type) cross for complex NP islands, they cross even more for subject islands, and are parallel (indicating absence of an island effect) for adjunct islands. Similar divergences from the human plots, when using the length-normalized PenLP score, also happen with GilBERTo (Figure A.3 in the Appendix vs Figure 5.3).

Conversely, for unidirectional models like GPT-2, the length-based normalization of the sentence score has a beneficial dampening effect to the model score numerical properties, as noted by Salazar et al. (2020). And this is also reflected in the plots for GePpeTto, with a divergence from human plots when using the non ideal, unnormalized, LP score (compare Figure 5.2 with Figure A.4 in the Appendix). However, we were not able to reproduce what described by Salazar et al. (2020) in their analysis of the numerical properties of unidirectional models, as compared to bidirectional ones. Specifically, our GPT-2 token surprisals plots do not show a descending curve (with high

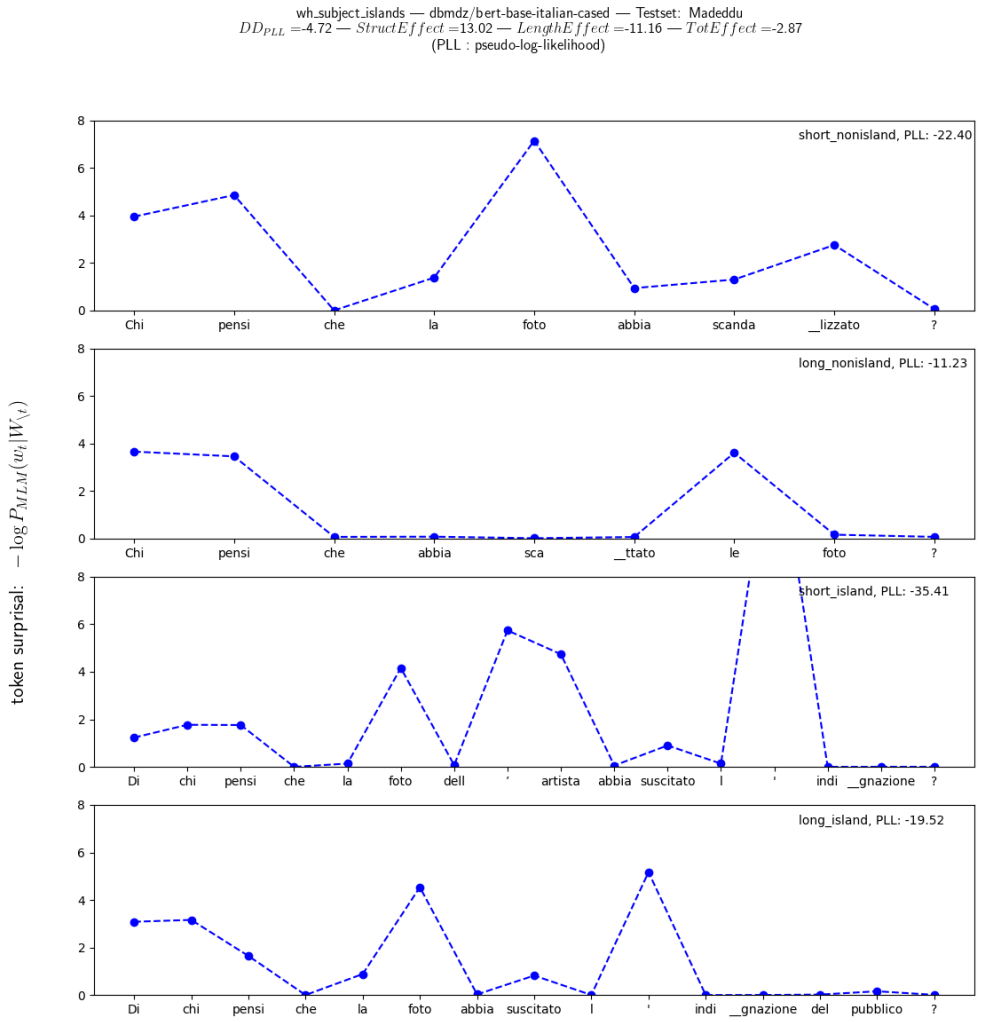


Figure 5.21: Token surprisal of an subject island item, as scored by the Italian BERT.

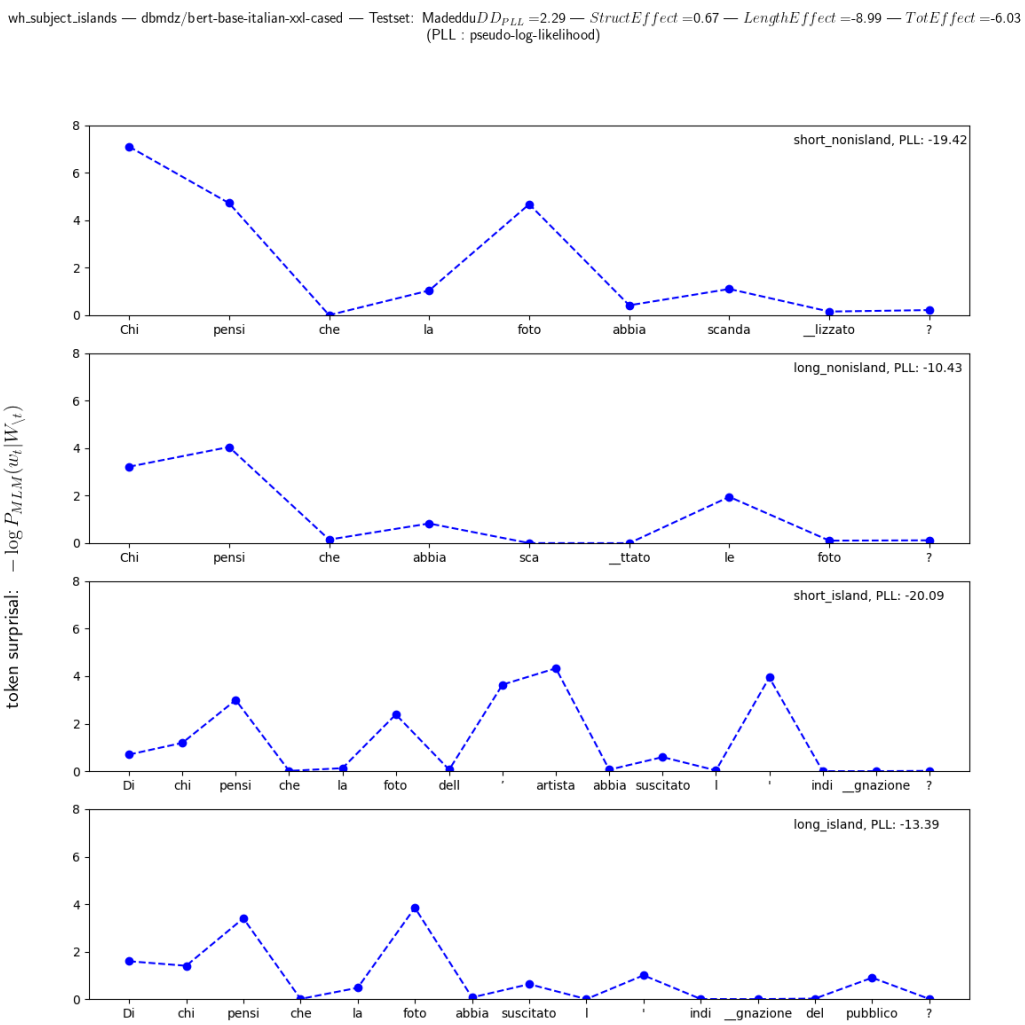


Figure 5.22: Token surprisal of an subject island item, as scored by the Italian BERT XXL.

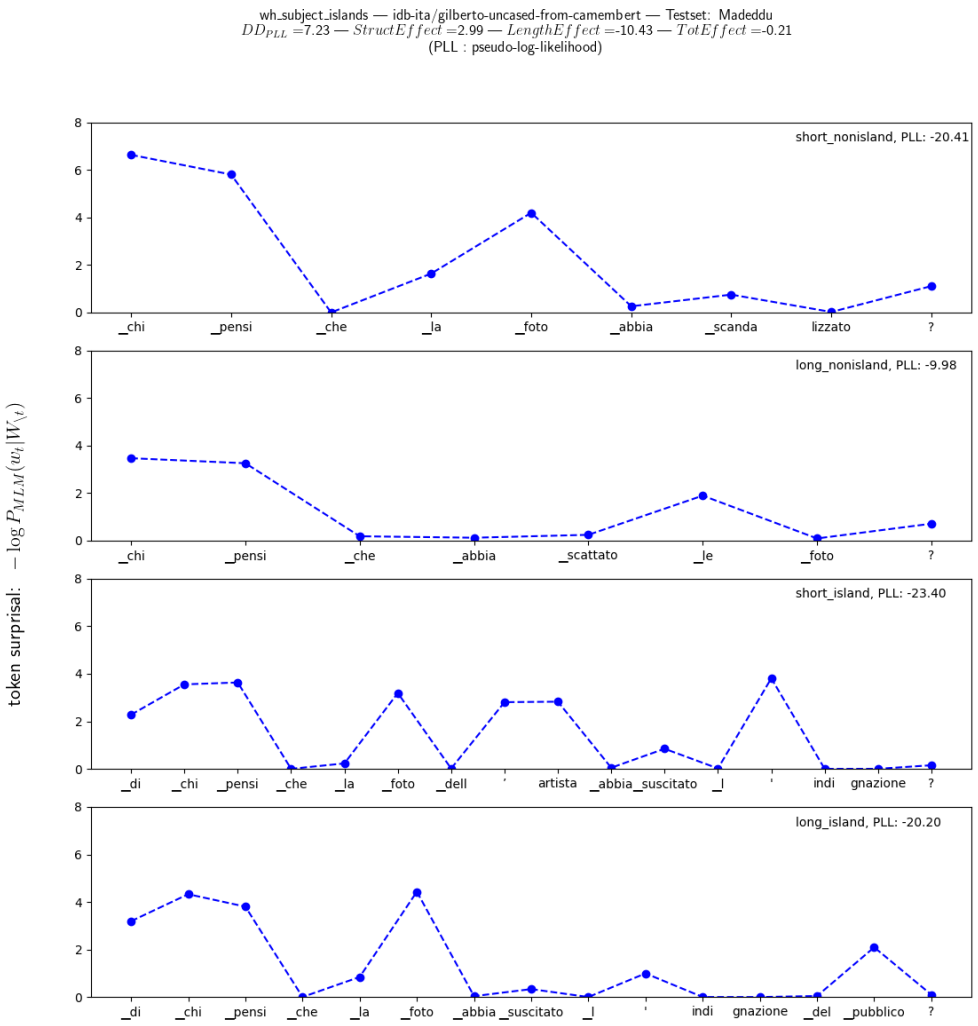


Figure 5.23: Token surprisal of an subject island item, as scored by GilBERTo.

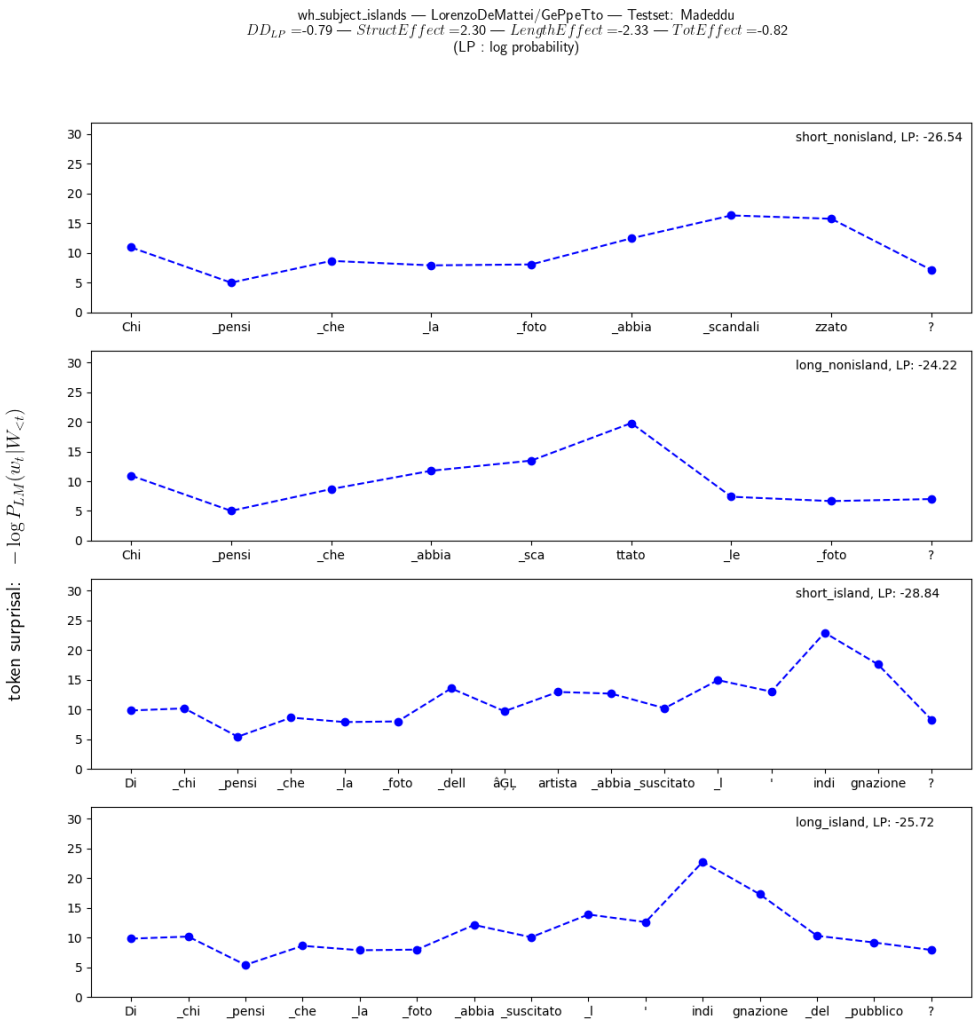


Figure 5.24: Token surprisal of a subject island item, as scored by GePpeTto.

surprisals for the first tokens in the sentence, that diminish as the model has access to more context). Instead, the trend in this plot seems to be also "flat" as those for the BERT-based models. This is possibly due to a methodological difference in the way we obtained the token surprisals from the GPT-2 output.

In general, we found that the token surpisal plots for GPT-2 to be much less clearly interpretable than those from the BERT-based models. One thing we were able to observe from the token surprisals from GPT-2, were surprisal spikes in correspondence of sub-units of OOV words.

In Figure 5.25 we observe the non-optimal sub-word units splits performed by the tokenizer used by GPT-2: the out-of-vocabulary (OOV) word *"impoverito"* (*"impoverished"*) has been split into *imp-ov-erito*, which clearly cannot take advantage of the semantically informative root *"pover-"* ("*poor*"). Bostrom and Durrett (2020) found that sub-word splitting algorithms like BPE are non optimal, and that a more linguistically plausible sub-word tokenization (with units that resemble more traditional morphemes), increases performance in downstream tasks. We argue that an optimal subword units tokenization should be a key element of a deep neural language models, since all the representations that these models learn are learned on top of these building blocks. We can expect that non optimal building blocks starting points will limit the possible generalizations and the patterns that these models are able to learn on top of them. The split of a training corpus in tokens is a strong signal that conveys a lot of information, and finding more optimal sub-word unit splits can also be though of as a form of regularization that avoids overfitting phenomena on the probability distribution of the training corpus.

5.3 Follow up experiments and future work

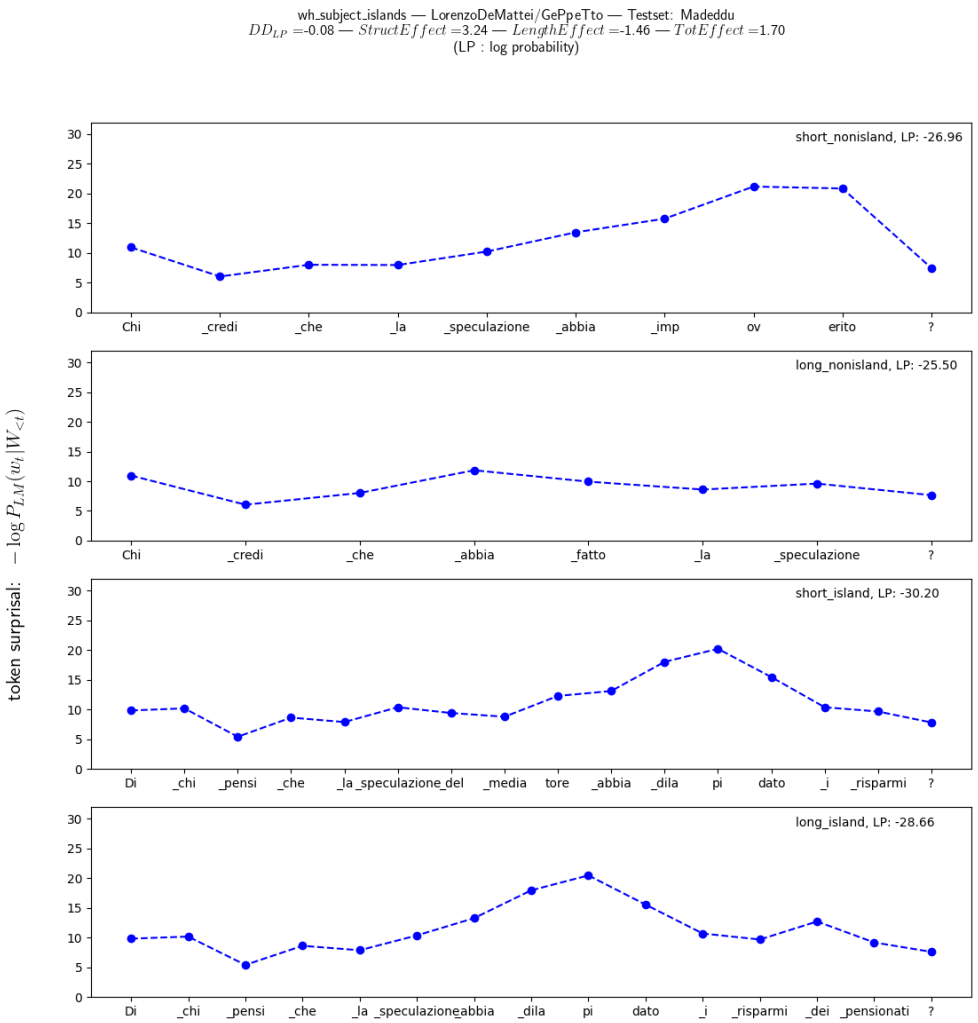


Figure 5.25: Token surprisal of a subject island item, as scored by GePpeTto.

Chapter 6

Conclusions

In assessing transformer-based language models, trained in the Italian language, on syntactic phenomena known as island effects, we found that the models responses for some types of island effects can mirror or resemble that of humans, as observed in previous psycholinguistic studies.

We found that the factorial experimental design we adopted from psycholinguistic studies was able to implicitly control for some unexpected confounds, when they were evenly distributed within the test items. However, as it is the case with assessment approaches based on minimal pairs, also in a factorial experimental setup sentences within an item should have ideally exactly the same lexical content, in order to avoid confounds. This is a challenging tasks in the development of tests covering more complex syntactic phenomena like island effects.

Our results confirm previous work that indicates that to master more complex syntactic phenomena, like island effects, modern transformer-based language models still require an amount of training data above the threshold of 100M tokens, which is the amount of data to which humans are exposed when they learn language, and also the threshold at which transformer-based models have been shown (although possibly by test suites not challenging enough) to acquire knowledge of most (but not all) syntactic and semantic phenomena.

In our contribution to the relatively novel line of research of the targeted syntactic evaluation of deep neural language models, an area for which adequate methodologies are sill in development, we found that it is an effective and informative approach to complement the performance scores achieved by the models, with the qualitative analysis of per-token surprisals, which reveal to which phenomena a model shows sensitivity. This combination of approaches seems also a promising avenue to reach better models explainability.

We agree with previous work (Wilcox et al., 2018; Ettinger, 2020) that observed that developing more comprehensive and more challenging targeted

linguistic tests for modern language model can play a significant role in accelerating their development.

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 Madeddu test suite

A.1.1 BERT

BERT XXL with the PenLP normalized sentence score

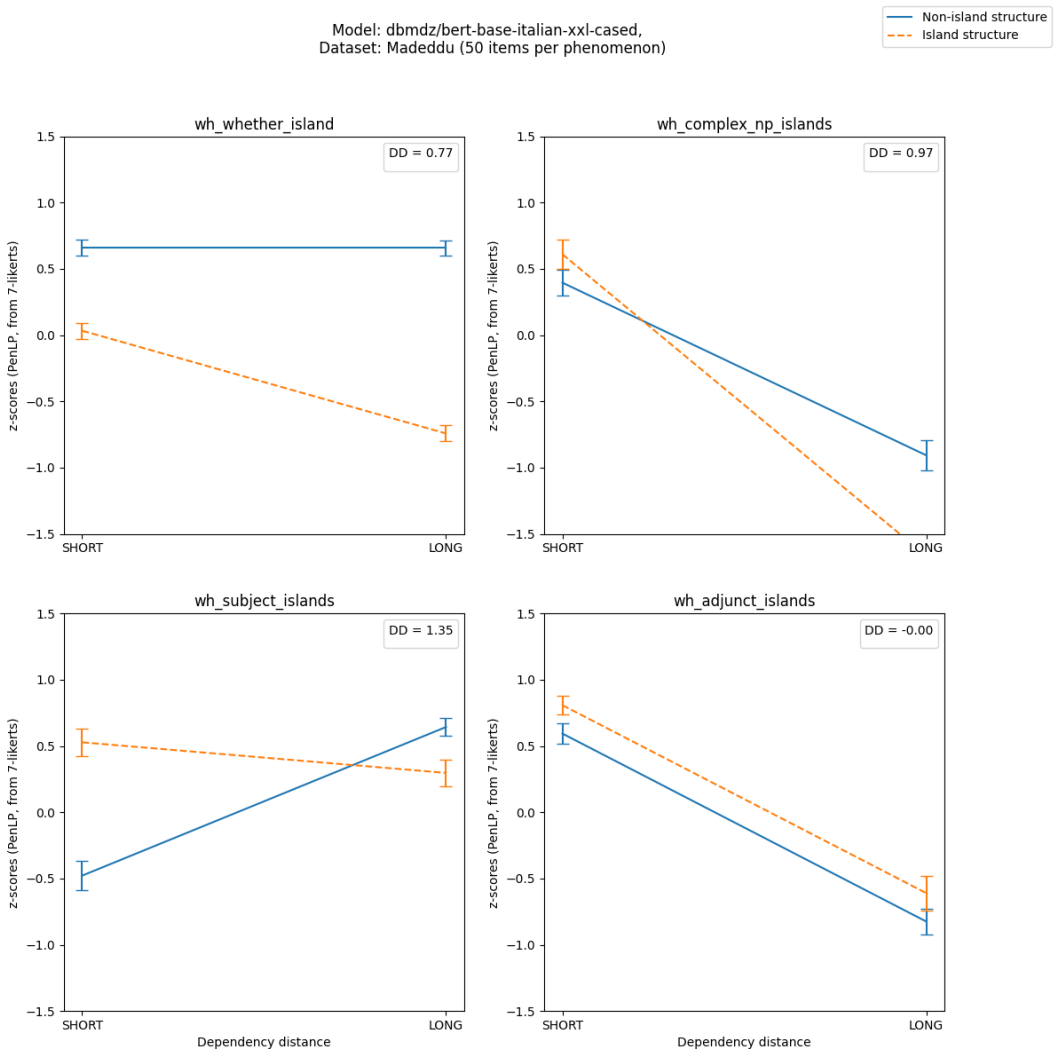


Figure A.1

BERT with the PenLP normalized sentence score

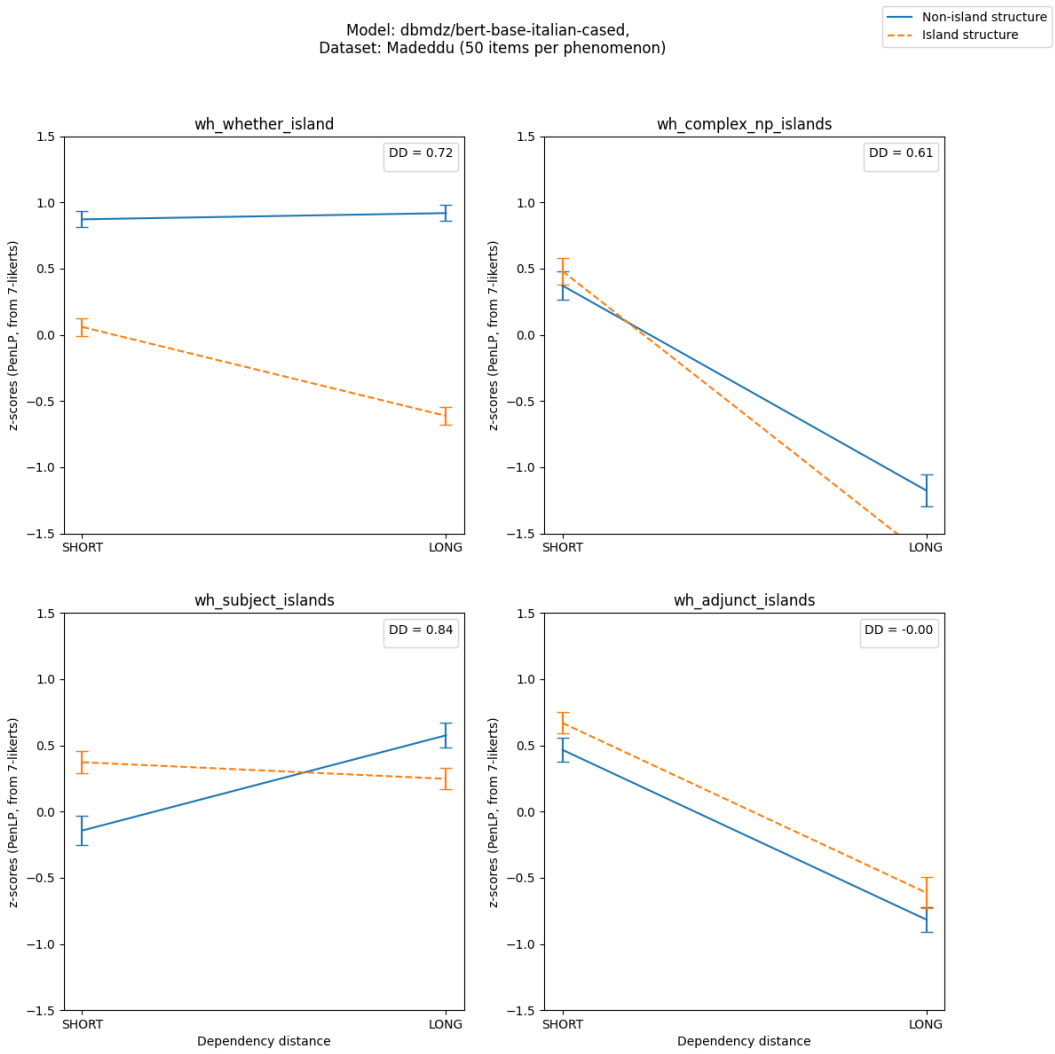


Figure A.2

A.1.2 GilBERTo

GilBERTo with the PenLP normalized sentence score

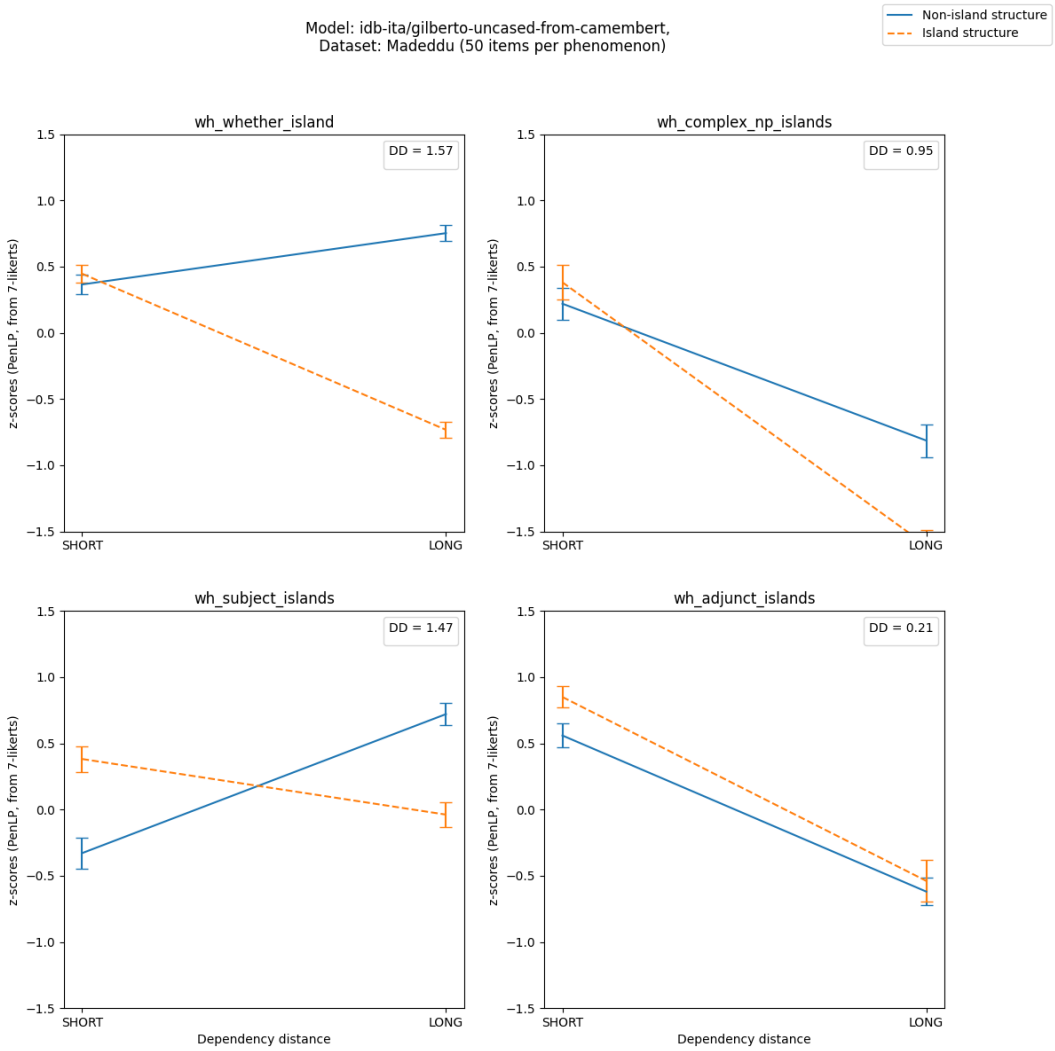


Figure A.3

A.1.3 GePpeTto

GePpeTto with the LP unnormalized sentence score

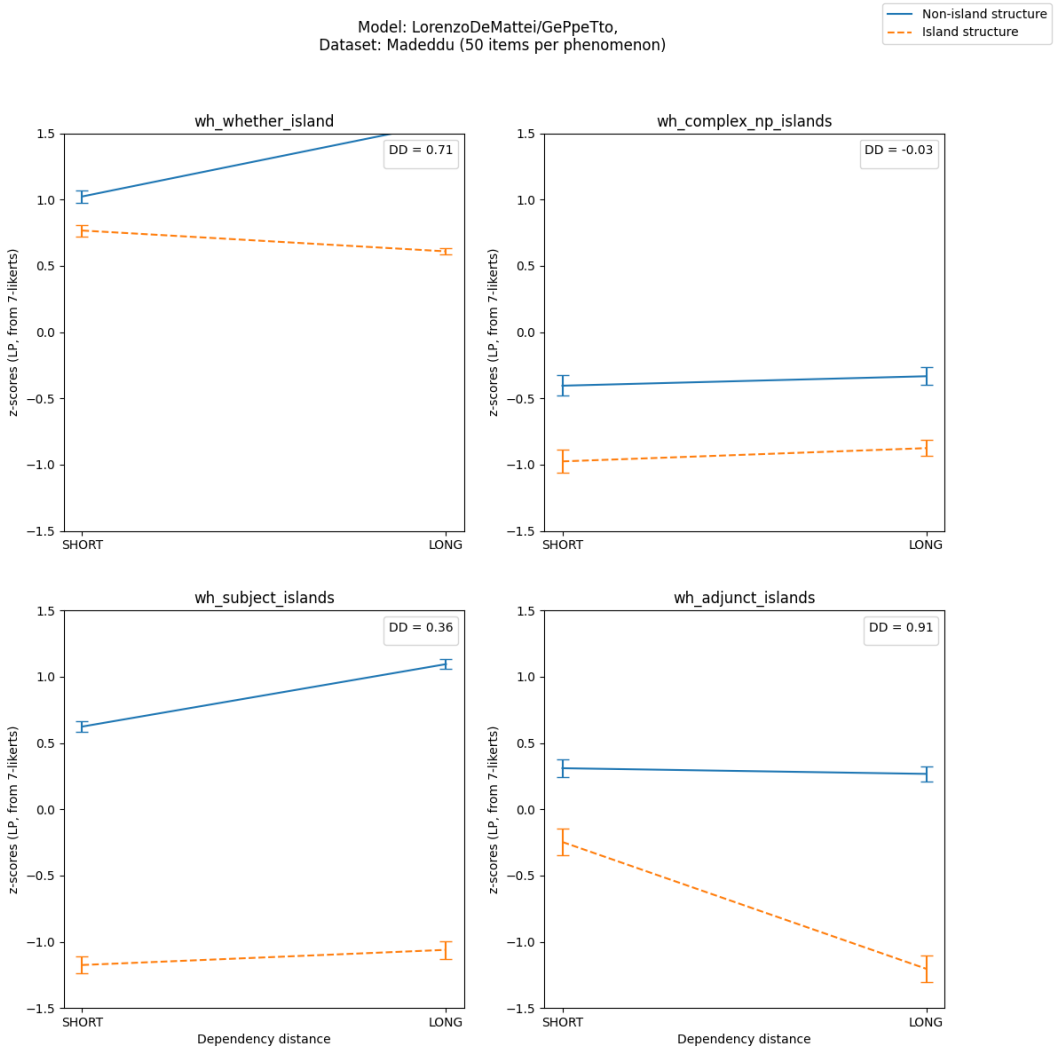


Figure A.4

A.2 Sprouse test suite

A.2.1 BERT

BERT XXL with the PenLP normalized sentence score

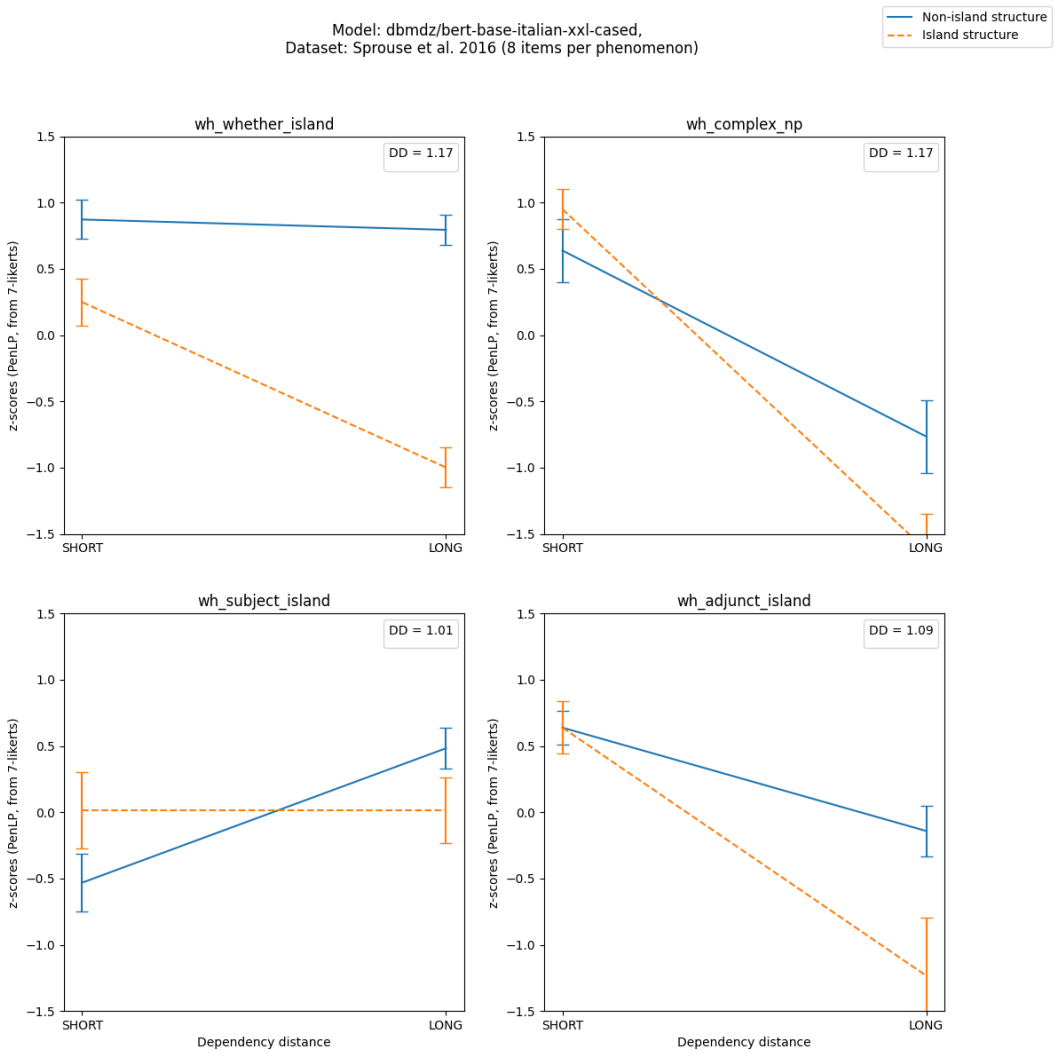


Figure A.5

BERT with the PenLP normalized sentence score

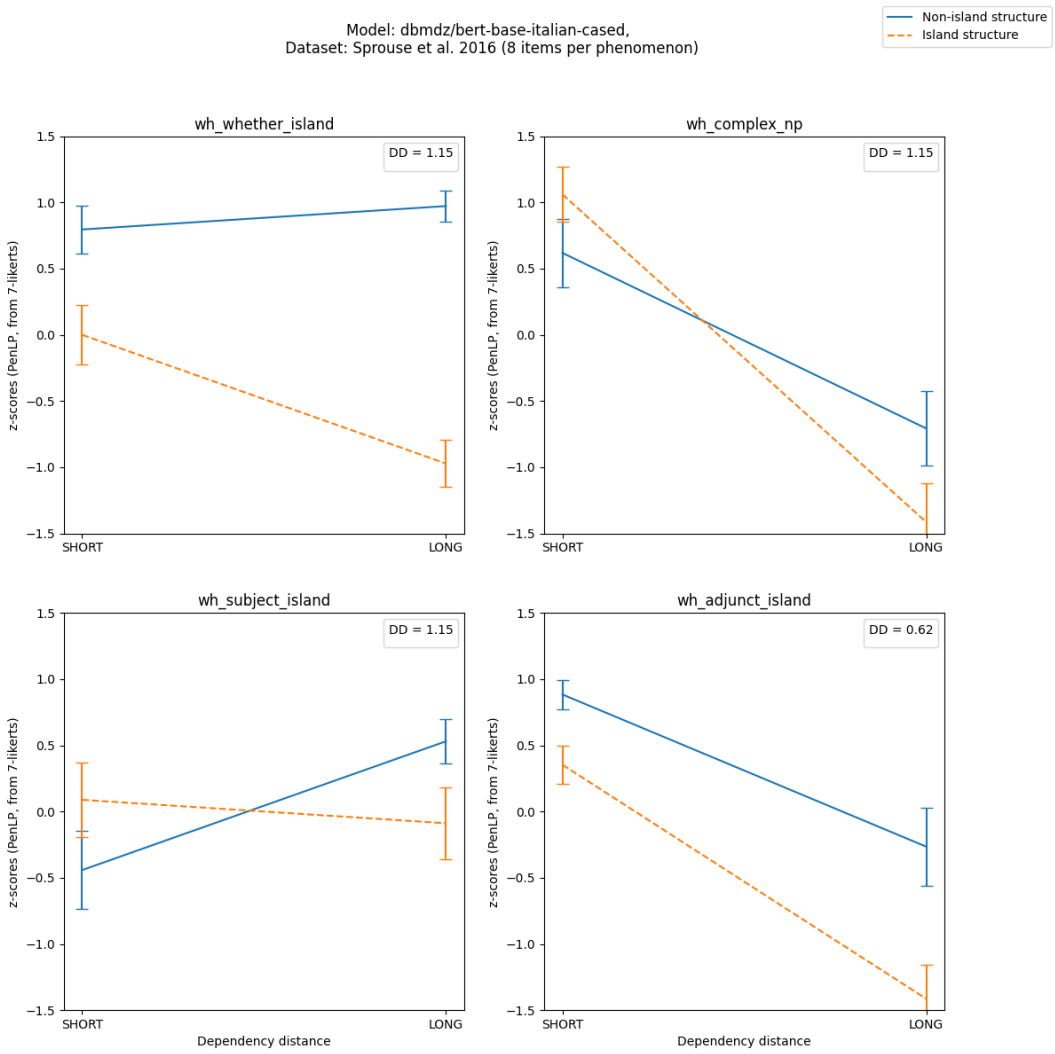


Figure A.6

A.2.2 GilBERTo

GilBERTo with the PenLP normalized sentence score

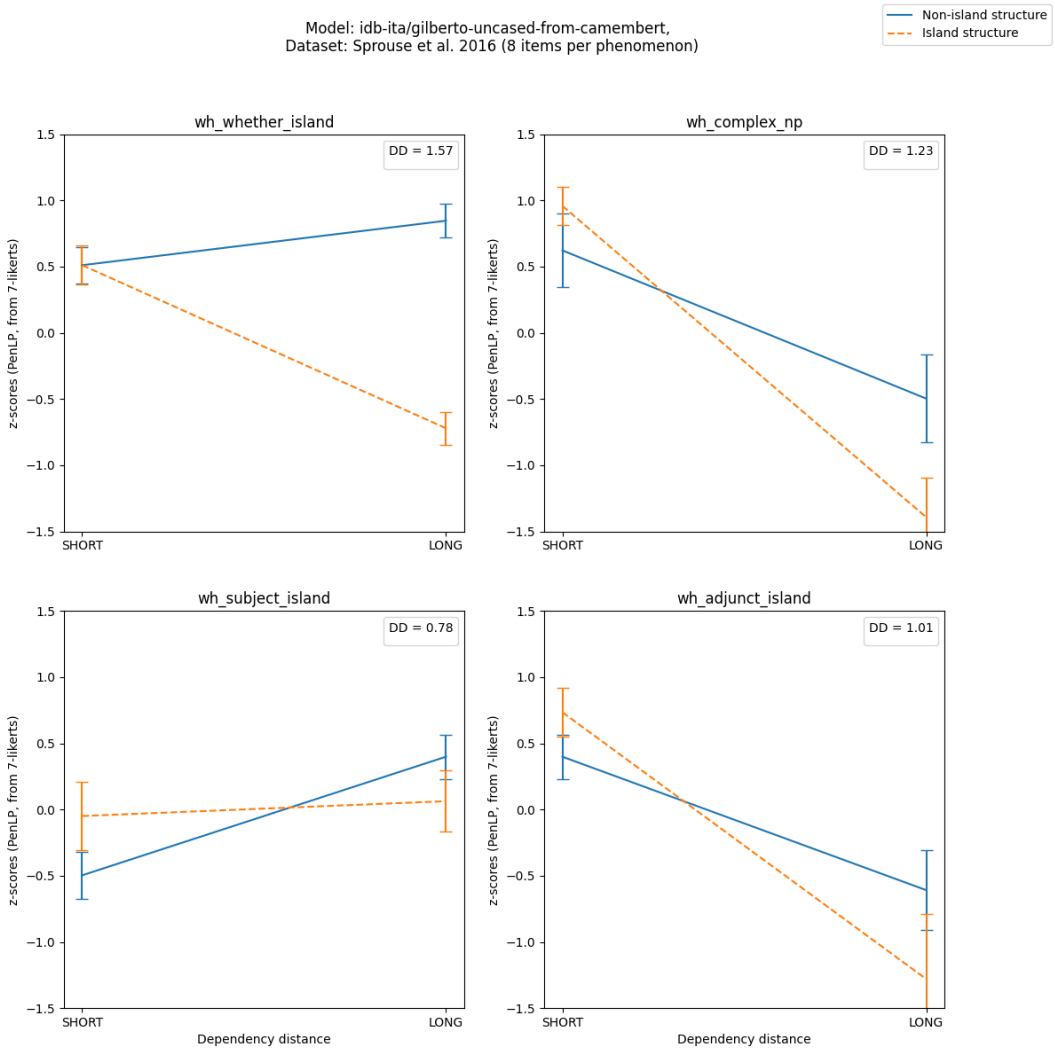


Figure A.7

A.2.3 GePpeTto

GePpeTto with the LP unnormalized sentence score

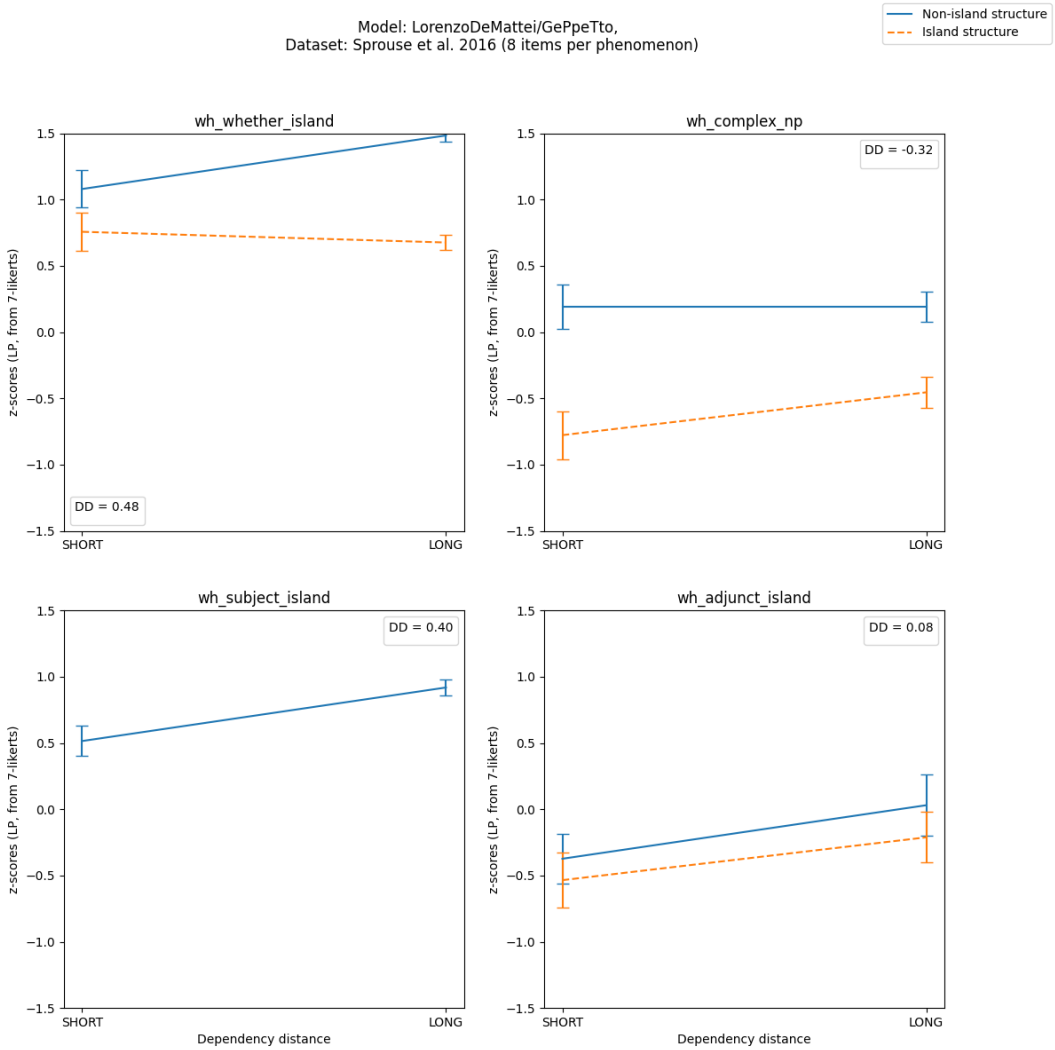


Figure A.8

A.3 Average factorial scores plots for the original test suites by Sprouse et al. (2016)

For comparison, in Fig.A.9-A.12 we also report the plots of the average factorial scores obtained by the models on the original test suites of the psycholinguistic study by Sprouse et al. (2016).

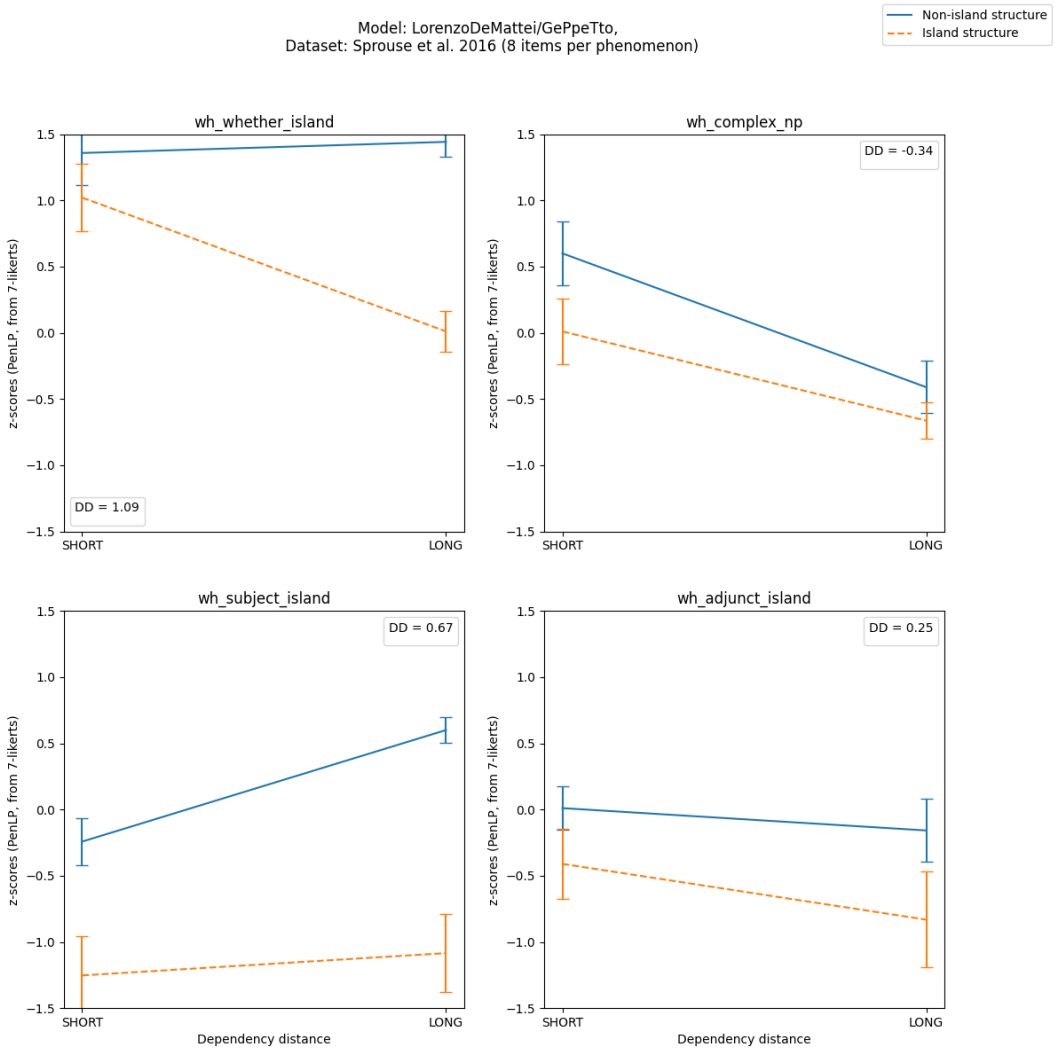


Figure A.9: Plots of average acceptability scores from GePpeTto, on the test suite by Sprouse et al. (2016).

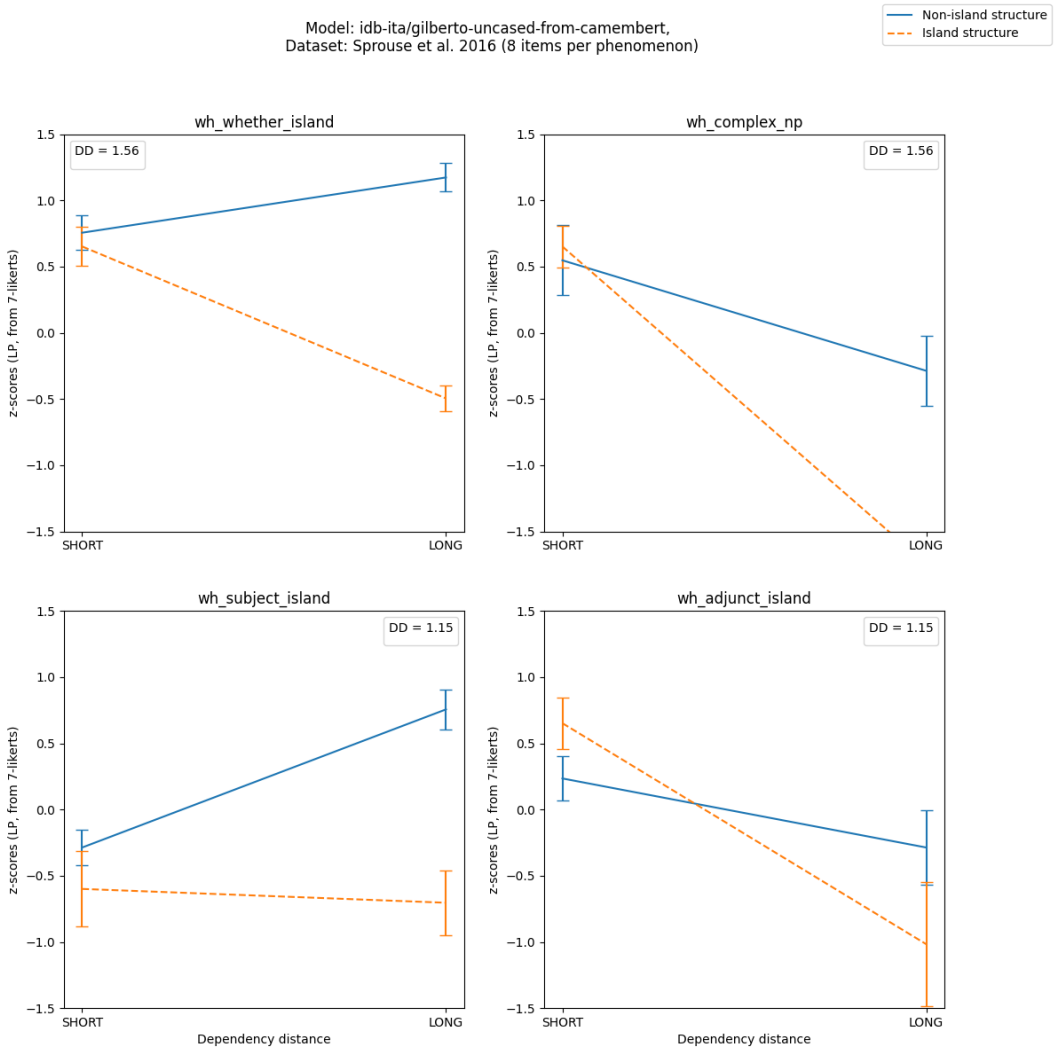


Figure A.10: Plots of average acceptability scores from GilBERTo, on the test suite by Sprouse et al. (2016).

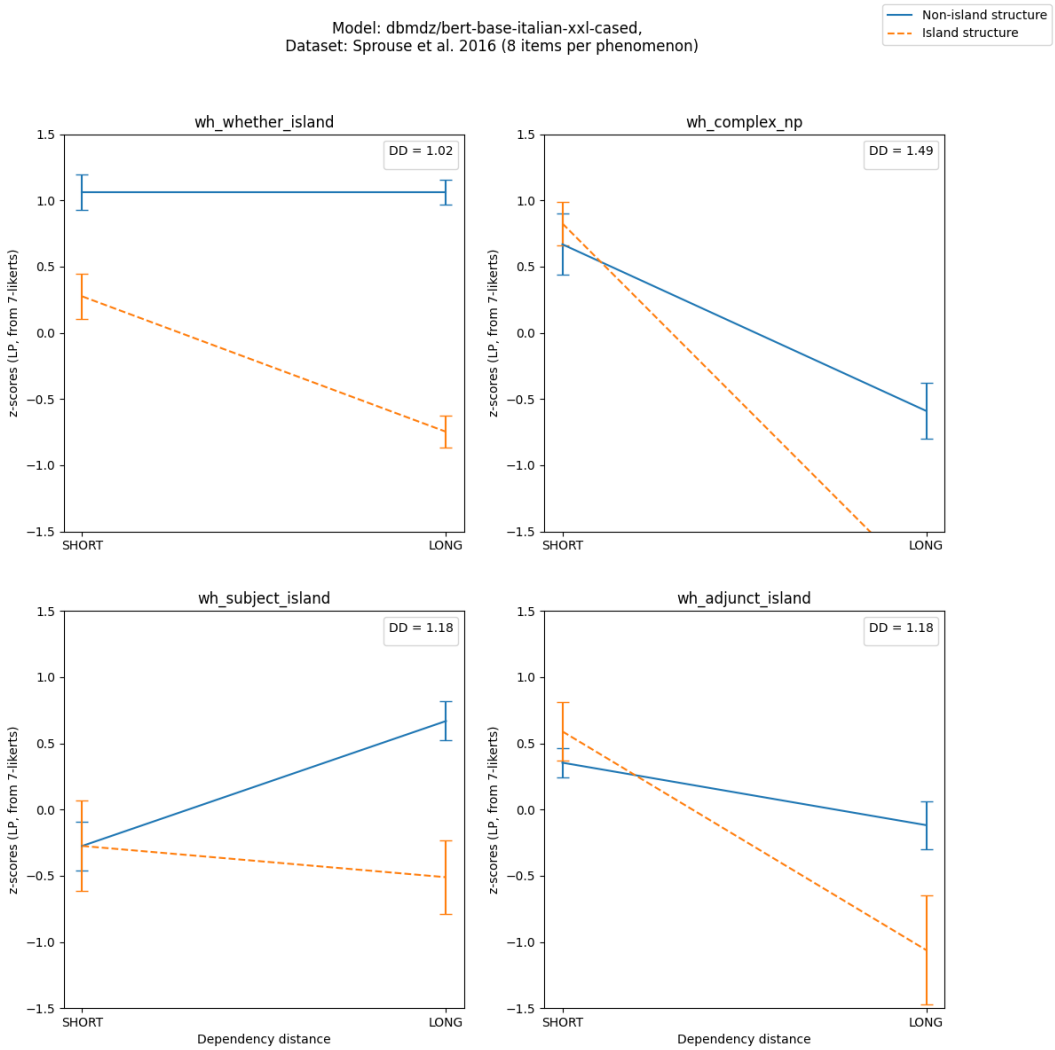


Figure A.11: Plots of average acceptability scores from BERT XXL (13B of training tokens), on the test suite by Sprouse et al. (2016).

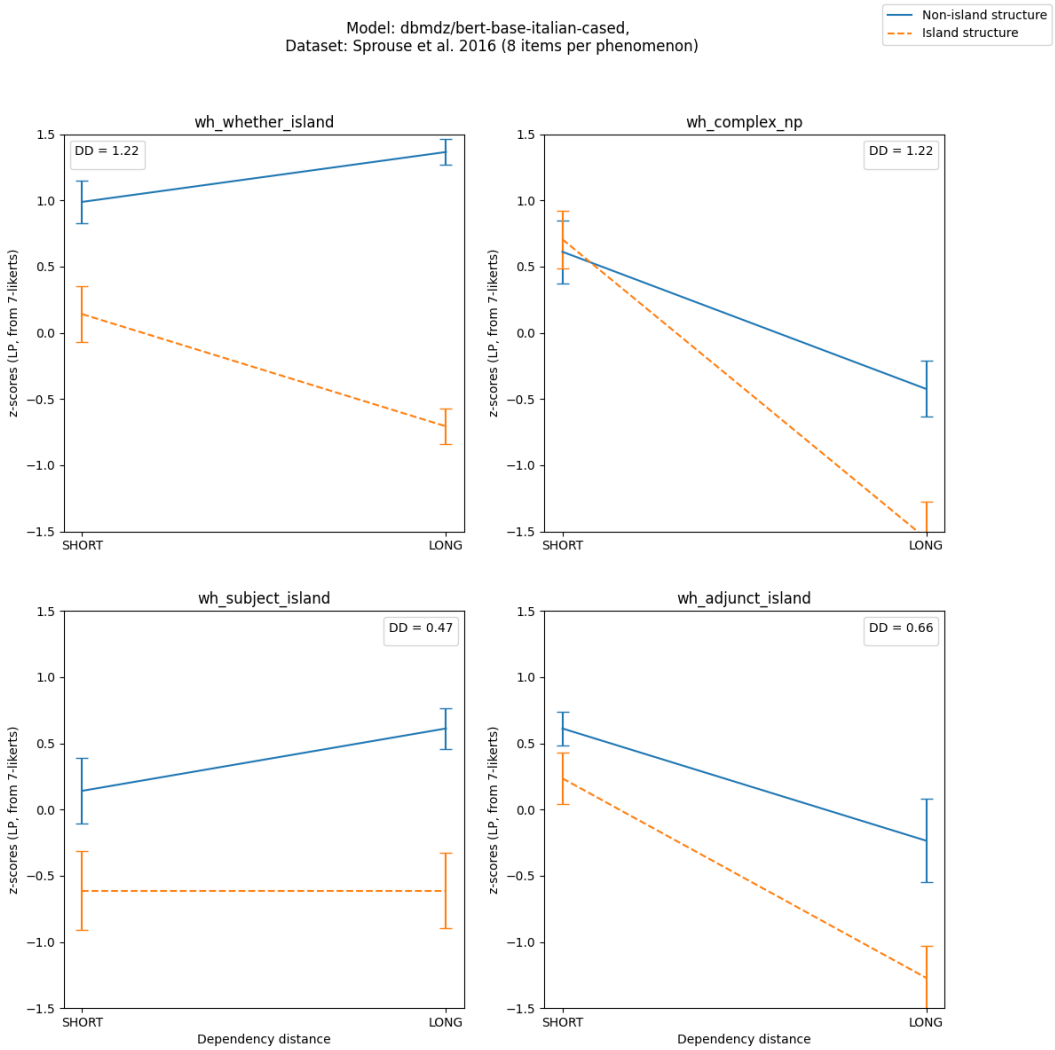


Figure A.12: Plots of average acceptability scores from BERT (2B of training tokens), on the test suite by Sprouse et al. (2016).

Bibliography

Ambridge, B. and Goldberg, A. E. (2008), 'The island status of clausal complements: Evidence in favor of an information structure explanation.', *Cognitive Linguistics* **19(3)**, 357–389.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003), 'A neural probabilistic language model', *Journal of Machine Learning Research* **3**, 1137–1155.

Bondevik, I., Kush, D. and Lohndal, T. (2021), 'Variation in adjunct islands: The case of norwegian', *Nordic Journal of Linguistics* **44**(3), 223–254.

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.

Cherniavskii, D., Tulchinskii, E., Mikhailov, V., Proskurina, I., Kushnareva, L., Artemova, E., Barannikov, S., Piontkovskaya, I., Piontkovski, D. and Burnaev, E. (2022), 'Acceptability judgements via examining the topology of attention maps', *arXiv preprint arXiv:2205.09630* .

Chomsky, N. (1957), 'Syntactic structures.'.

Chomsky, N. (1965), *Aspects of the theory of syntax*, The MIT Press.

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.

Cowart, W. (1997), *Experimental syntax: Applying objective methods to sentence judgments*, Sage Publications, Thousand Oaks.

Da, J. and Kasai, J. (2019), Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations, *in* 'Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing', pp. 1–12.

Dąbrowska, E. (2015), 'What exactly is universal grammar, and has anyone seen it?', *Frontiers in psychology* **6**, 852.

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.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Downing, A. and Locke, P. (2002), *A University Course in English Grammar*, Psychology Press.

Elazar, Y., Ravfogel, S., Jacovi, A. and Goldberg, Y. (2021), 'Amnesic probing: Behavioral explanation with amnesic counterfactuals', *Transactions of the Association for Computational Linguistics* **9**, 160–175.

Ettinger, A. (2020), 'What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models', *Transactions of the Association for Computational Linguistics* **8**, 34–48.

Fodor, J. D. (1978), 'Parsing strategies and constraints on transformations', *Linguistic Inquiry* **9**(3), 427–473.

Forbes, M., Holtzman, A. and Choi, Y. (2019), 'Do neural language representations learn physical commonsense?', *arXiv preprint arXiv:1908.02899* .

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M. and Levy, R. (2019), Neural language models as psycholinguistic subjects: Representations of syntactic state, *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)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 32–42.
URL: *https://aclanthology.org/N19-1004*

Gulordava, K., Bojanowski, P., Grave, É., Linzen, T. and Baroni, M. (2018), Colorless green recurrent networks dream hierarchically, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', pp. 1195–1205.

Hauser, M. D., Chomsky, N. and Fitch, W. T. (2002), 'The faculty of language: what is it, who has it, and how did it evolve?', *Science* **298**(5598), 1569–1579.

Hawkins, J. A. (1999), 'Processing complexity and filler-gap dependencies across grammars', *Language* **75**(2), 244–285.

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.

Hofmeister, P. and Sag, I. A. (2010), 'Cognitive constraints and island effects', *Language* **86**(2), 366.

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/.

Khalid, U., Beg, M. O. and Arshad, M. U. (2021), 'Rubert: A bilingual roman urdu bert using cross lingual transfer learning', *arXiv preprint arXiv:2102.11278* .

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. (2016), 'Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge', *Cognitive science* **41**(5), 1202–1241.

Linzen, T. (2018), 'What can linguistics and deep learning contribute to each other?', *arXiv preprint arXiv:1809.04179* .

Linzen, T. (2020), How can we accelerate progress towards human-like linguistic generalization?, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 5210–5217.

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, N. F., Gardner, M., Belinkov, Y., Peters, M. E. and Smith, N. A. (2019), Linguistic knowledge and transferability of contextual 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. 1073–1094.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692* .

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.

Maxwell, S. E. and Delaney, H. D. (2003), *Designing experiments and analyzing data: A model comparison perspective*, Mahwah: Lawrence Erlbaum Associates.

Newman, B., Ang, K.-S., Gong, J. and Hewitt, J. (2021), Refining targeted syntactic evaluation of language models, *in* 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', pp. 3710–3723.

Nikoulina, V., Tezekbayev, M., Kozhakhmet, N., Babazhanova, M., Gallé, M. and Assylbekov, Z. (2022), 'The rediscovery hypothesis:: Language models need to meet linguistics', *Journal of Artificial Intelligence Research* **72**, 1343–1384.

Pearl, L. and Sprouse, J. (2013), 'Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem', *Language Acquisition* **20**(1), 23–68.

Pullum, G. K. and Scholz, B. C. (2002), 'Empirical assessment of stimulus poverty arguments', *The linguistic review* **19**(1-2), 9–50.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (n.d.), 'Language models are unsupervised multitask learners'.

Ribeiro, M. T., Wu, T., Guestrin, C. and Singh, S. (2020), Beyond accuracy: Behavioral testing of nlp models with checklist, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 4902–4912.

Rizzi, L. (1982), 'Violations of the wh-island constraint and the subjacency condition', *Issues in Italian syntax* pp. 49–76.

Rogers, A., Kovaleva, O. and Rumshisky, A. (2020), 'A primer in bertology: What we know about how bert works', *Transactions of the Association for Computational Linguistics* **8**, 842–866.

Ross, J. R. (1967), 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.

Sorace, A. and Keller, F. (2005), 'Gradience in linguistic data', *Lingua* **115**(11), 1497–1524.

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.

Sprouse, J. and Hornstein, N. (2013), 'Experimental syntax and island effects: Toward a comprehensive theory of islands', *Experimental syntax and island effects* pp. 1–20.

Sprouse, J., Wagers, M. and Phillips, C. (2012), 'A test of the relation between working-memory capacity and syntactic island effects', *Language* pp. 82–123.

Szabolcsi, A. (2006), 'Strong vs. weak islands', *The Blackwell companion to syntax* **4**, 479–531.

Tenney, I., Das, D. and Pavlick, E. (2019), Bert rediscovers the classical nlp pipeline, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', pp. 4593–4601.

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**.

Wallace, E., Wang, Y., Li, S., Singh, S. and Gardner, M. (2019), Do nlp models know numbers? probing numeracy in 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)', pp. 5307–5315.

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.

Wu, Z., Peng, H. and Smith, N. A. (2021), 'Infusing finetuning with semantic dependencies', *Transactions of the Association for Computational Linguistics* **9**, 226–242.

Zhang, Y., Warstadt, A., Li, X. and Bowman, S. (2021), When do you need billions of words of pretraining data?, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', pp. 1112–1125.