La non canonica l'hai studiata? Exploring LLMs and Sentence Canonicity in Italian

Claudiu Daniel Hromei[1,*], Danilo Croce[1], Rodolfo Delmonte[2] and Roberto Basili[1]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
[2]Ca' Foscari University, Venice, Italy

Abstract

This paper investigates the ability of Large Language Models (LLMs) to differentiate between canonical and non-canonical sentences in Italian, employing advanced neural architectures like LLaMA and its adaptations. Canonical sentences adhere to the standard Subject-Verb-Object (SVO) structure. We hypothesize that recent generative LLMs are influenced heavily by the English language, where non-canonical structures are very rare. Using the in-context learning technique, we probe these models and further fine-tune them for this specific task. Initial results indicate that these models continue to struggle with this task even after fine-tuning. Additionally, we introduce a test set comprising several hundred sentences from the poetry domain, which presents significant challenges for the canonical structure task.

Keywords

Large Language Models, Italian Sentence Structure, Non-Canonical Structures, In-Context Learning

1. Introduction

Unlike contemporary English, which primarily follows a Subject-Verb-Object (SVO) structure, Italian exhibits a rich variety of non-canonical syntactic structures that deviate from this pattern[1] [1, 2]. Italian is generally considered a configurational language with a neutral or canonical SVO sentence structure. However, it also displays characteristics of a weak non-configurational language due to several typological parameters: free subject inversion, pro-drop, and nonlexical expletives. Additionally, Italian lacks wh- in situ, preposition stranding, deletable complementizers, impersonal passives, and parasitic gaps with the same argument [3].

In cognitive linguistic terms, the use of surface or syntactic constituency and word order in non-canonical sentences in Italian reflects its informational structure. As an example, the first sentence "*Sempre caro mi fu quest'ermo colle e questa siepe che da tanta parte de l'ultimo orizzonte il guardo esclude*"[2] of Leopardi's famous *L'infinito* is a typical example of a non-canonical sentence: the complement is fronted and the subject is in post-verbal position, also known as complete argument inversion, for letting the reader focus on the subject and main verb rather than the complement.

The functional or relational interpretation of these syntactic structures, along with semantic processing, is essential to understanding the semantic roles associated with displaced grammatical functions. For instance, when a subject appears in an inverted position, it indicates a pragmatically motivated displacement, emphasizing focus over an otherwise topic-related function. Typically, subjects, understood as topics or "what the sentence is about" and constituting old information, precede the verb. This is consistent with Italian and English, both of which follow an SVO structure. Conversely, focus, defined as "the essential piece of new information carried by a sentence," usually follows the verb in the "comment" portion of the sentence.

We consider complexity measures sensitive to non-canonical structures (NCS), which are pragmatically motivated and used to encode structured meaning with high informational content, related to the FOCUS/TOPIC non-argument functions in Lexical-Functional Grammar (LFG) [4, 5]. Non-canonical structures can aid the reader or interlocutor in better understanding the pragmatically relevant meaning in context [6].

Italian NCS are relatively frequent in text. In [7], the authors analyzed the VIT (Venice Italian Treebank) by manually annotating non-canonical structures and inflected propositions in Italian. The study found that Left Dislocated Complements, where a complement of the main verb according to subcategorization restrictions occurs, appear in 0.03% of cases. Dislocated Subjects, indicating any NP subject not followed by the main verb, occur in 0.28% of cases. The overall percentage of non-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
hromei@ing.uniroma2.it (C. D. Hromei); croce@info.uniroma2.it (D. Croce); delmont@unive.it (R. Delmonte); basili@info.uniroma2.it (R. Basili)
0009-0000-8204-5023 (C. D. Hromei); 0000-0001-9111-1950 (D. Croce); 0000-0003-0282-7661 (R. Delmonte); 0000-0001-5140-0694 (R. Basili)

[1]Elizabethan English was more similar to Italian in its variety of syntactic structures.
[2]In English: "*Always dear to me was this solitary hill and this hedge which from large side of the ultimate horizon the gaze excludes*"

projectivity in written texts is 7%, based on 230,629 constituents. Compared to Latin, where the non-projectivity index is 6.65% in the Latin Dependency Treebank containing about 55,000 tokens, Italian and Latin are quite similar. In contrast, English tree projectivity in the Penn Treebank (PT), where the majority of data corresponds to the articles of Wall Street Journal (WSJ), shows much lower numbers: with 720,086 constituents, the non-projectivity index is 0.01004%.

Thus, Italian speakers have high expectancies for the presence of an NCS due to processing difficulties also raised by the number of unexpressed subjects: 61% of all Inflected Propositions lack a lexically expressed subject. This does not apply to English speakers, for whom NCS are infrequent and context-specific. In this view, Italian is considered unique for its use of many of the non-canonical structures found in contemporary poetry and examined in this experiment. The richness and freedom of the language give the speakers the ability to produce such a diverse typology of non-canonical structures, which stems from its Latin heritage, with the Null Subject being one of the most well-known features. Like many other languages, including Spanish, Portuguese, and Catalan, as well as Chinese, Japanese, Slavic languages, Greek, and Hebrew, Italian is a Null Subject Language. However, this parameter alone does not fully explain the richness and complexity of syntactic structures seen in Italian poetry. While other Romance languages share similar syntactic traits, the specific linguistic legacy and poetic traditions of Italian give it a unique character in this regard.

In this paper, we want to analyze the ability of recently proposed Large Language Models to detect non-canonical sentences in Italian. Our hypothesis is that, given the very large percentage of English training data (usually more than 90%) and the very low percentage of Italian training data (usually less than 1%), these models have a limited capacity to process such structures and they rely mostly on the English writing structures. On the other hand, the models that have been specifically adapted or fine-tuned on Italian data should show a better understanding of the canonicity in Italian.

In the rest, Section 2 describes the related work, Section 3 shows the approach in recognizing the canonical structures, Section 4 presents and discusses the results, while Section 5 derives the conclusions.

2. Related Work

Our approach has been previously adopted by other researchers but with slightly different aims, as described below. Initial attempts at parsing Italian treebanks of constituent structures focused on two small treebanks: TUT [8, 9] and ISST [10], containing approximately 3,500 and 3,000 sentences, respectively. *Illo tempore*, these efforts yielded an F1 score of 82.96%, while comparable parsers (Stanford, Collins, and MaltParser) achieved about 92.10% on the WSJ treebank. The lower performance in Italian was primarily due to two factors: a higher number of non-canonical structures (i.e., word order variations) and the presence of pro-drop clauses, where the subject is lexically omitted — a challenge also documented for other similar languages [11].

Significant improvements in parsing performance were noted in a paper on the EVALITA shared task on constituency parsing, where the best F1 score increased from 70% to 84% [12], attributed to the nearly doubling of training samples between 2007 and 2011. In [13], the authors presented a new dataset of Italian based on "marked" sentences to test the performance of the neural parser TINT. The result for LAS dependency structures was 77% accuracy, three points below the best results on the UD corpus of Italian, which was 80%. This outcome confirmed previous findings with a small dataset of strongly marked sentences, where accuracy was below 50%. The authors detailed seven types of marked structures in their treebank corpus: cleft, left-dislocated, right-dislocated, presentative "*ci*" (*there* in English), inverted subject, pseudo-clefts, and hanging topic, with cleft and left-dislocated sentences being the most common.

In this context, it is interesting to explore the capabilities of state-of-the-art methods for addressing the problem of distinguishing between canonical and non-canonical sentences in Italian. This exploration is motivated by the complexity and richness of Italian syntax, which presents unique challenges for natural language processing models. Mostly all actual state-of-the-art models are based on the Transformer architecture [14]. This game-changer model comprises two main components, leading to different model families. The encoder, used in models like BERT [15], RoBERTa [16], and Sentence BERT [17], encodes input sequences using self-attention. In contrast, decoders, such as GPT [18], GPT-3 [19], and LLaMA [20], generate output sequences auto-regressively. Beyond these, encoder-decoder models like T5 [21] and BART [22] integrate both components, excelling in tasks such as translation, summarization, and question-answering.

One notable Transformer-based architecture is the LLaMA foundational model [20]. LLaMA is a large model with billions of parameters that generates output sequences auto-regressively based on the input and previously generated tokens. It has been recently applied to a variety of linguistic tasks by instruction-tuning a monolithic architecture to solve them all [23]. This family of models is promising as they rely on auto-regressive generation methods and, thanks to their massive amount of training data and parameters, can solve a plethora of

linguistic tasks. Additionally, [24] demonstrated the application of LLaMA-family models for syntactic parsing across multiple languages, highlighting the capability of the model to analyze and detect sentence structures. This work underscores the versatility of large language models in handling diverse syntactic frameworks, further probing their performance in cross-linguistic scenarios. Finally, architectures specifically adapted for Italian, such as Camoscio [25] and LLaMAntino [26], are tuned with instruction datasets for the Italian language, starting from the original LLaMA model and its second variant, LLaMA2-chat, respectively. They demonstrate a strong understanding of the language and an excellent ability to generate appropriate responses.

In this paper, we aim to explore the ability of Large Language Models (LLMs) to distinguish between canonical and non-canonical sentences in Italian using neural architectures such as LLaMA and its various adaptations, as discussed in the next Section. It's interesting to note that in the future one might explore the applications of probing syntax at the intermediate layers of various models.

3. Recognizing Canonical structures through LLMs

To address the capabilities of Large Language Models in recognizing the canonical structures, they can be utilized through In-Context Learning techniques [27] or by directly fine-tuning the model for specific downstream tasks. In-context learning relies on the model's pre-existing knowledge acquired during pre-training and on instructions provided in natural language at inference time. This method does not involve additional training and can be categorized based on the number of examples provided: *i)* **0-shot Learning**, where no examples are given, and the model generates responses based solely on its pre-existing knowledge and the provided instructions; *ii)* **1-shot Learning**, where one example per class (positive and negative in our case) is added to provide a more precise context, these examples help the model better understand the task by offering a concrete reference point; *iii)* **Few-shot Learning**, where more than one example per class is provided to give the model additional contextual information during decision-making. This approach is particularly effective when very few examples (such as 2 or 4) are given, but it can be extended up to the maximum input context length.

For both one-shot and few-shot learning approaches, a key challenge is selecting the most informative examples to provide during inference. One effective strategy is to retrieve examples that are most similar to the current sequence to be classified, focusing on those with a similar structure or meaning. A commonly used method for this is to generate vector embeddings of sentences using a model like sBERT [17]. This model produces a contextualized vector that represents the information contained in a sentence. By applying Cosine Similarity, we can rank these vectors and select the training examples most similar to the input sequence. This process ensures that the model is supplied with the most relevant solved examples for a given input. It's important to note that these examples may not always capture the same explicit syntax representation as a Tree Kernel [28] function would, in which every word of the sentence is explicitly annotated with syntactic information and linked to each other. However, the crucial aspect is that the examples provided are sufficiently similar in meaning and context, and the sBERT architecture is very effective.

When the model's pre-existing knowledge is insufficient, we can fine-tune it on the downstream task. Fine-tuning involves training the model in a traditional manner using input-output pairs (training data) to adjust its parameters. This process improves the model's performance on specific tasks, allowing it to learn from a more extensive set of examples. As a result, the model becomes more adept at handling similar queries in the future, with a focus on the specific task at hand. By leveraging these techniques, LLMs can recognize and respond to canonical structures with varying degrees of efficiency and accuracy.

3.1. Training LLMs against non-Canonical structures

To interact with the models, we need a sufficiently detailed prompt, which includes a natural language description of the task (i.e., the rules to determine whether an Italian sentence follows the canonical structure) and specifies the type of answer we expect the LLM to produce: *Sì* (*Yes* in English) if the sentence is canonical and follows the rules, or *No* otherwise. For the training and the 0-shot strategy, we used the following prompt:

> *"Dimmi se la seguente frase ha una struttura canonica o meno. Per Canonica si intende una frase che segue una struttura standard per ogni verbo presente. Più nello specifico, le frasi canoniche seguono queste regole: contengono SOLO sequenze del tipo nome o strutture nominali SEGUITE da struttura verbale a sua volta seguita (oppure no) da complementi OPPURE contengono SOLO sequenze composte da struttura verbale seguita da complementi, dove: STRUTTURE VERBALI sono sequenze composte da ausiliare o/e modale e verbo, e tra i due ci può essere un avverbio oppure strutture preposizionali COMPLEMENTI sono strutture nominali oppure strutture preposizionali*

oppure strutture frasali oppure strutture infinitivali. Tutte le altre frasi sono da considerarsi come Non Canoniche. Riguardo il prossimo input, rispondi 'sì' se è 'canonico', 'no' se è 'non canonico'."

For the 1-shot scenario, immediately after the above prompt, we append the following instruction, where the two provided examples are selected as the most relevant for the input example:

> Ti faccio un paio di esempi:
> <Positive_Example> e devi rispondere sì.
> <Negative_Example> e devi rispondere no.

When fine-tuning a model, a highly detailed prompt might seem excessive, especially since traditional training involves repeating the prompt multiple times. However, our hypothesis is that clearly explaining the task to the model aids in faster convergence of the parameters and a more rapid reduction in loss during training. Therefore, this is the reason why our prompt includes a comprehensive description of the canonical sentence structure. This description details that each verb must adhere to specified constraints, the types of sequences they can contain, the verbal structures, and the order of complements. If a verb does not adhere to these constraints, it should be classified as Non-Canonical.

3.2. LLM architectures of non-Canonical structures

Today, the landscape of Large Language Models (LLMs) is vast, making it challenging to choose the most suitable model. In this paper, we focus on several well-known models from the LLaMA family: LLaMA1 [20], the first in the series; LLaMA2 [29], which introduced minor improvements in Transformer architecture; Camoscio [25], an instruction-tuned LLaMA model fine-tuned on Italian data; ExtremITA [23], an architecture designed for a wide range of Italian tasks; and LLaMAntino [26], an adaptation of the original LLaMA2 model for the Italian language.

We expect the best-performing models to be those specifically adapted or fine-tuned on Italian data, such as Camoscio, ExtremITA, or LLaMAntino. One significant issue with the English models is that non-canonicity is very rare in English, as the language predominantly follows the Subject-Verb-Object structure, which is canonical, with very few (grammatically correct) non-canonical examples.

4. Empirical Investigation

In this setup, the models trained and those utilized in the k-shot scenario are required to answer Yes if the given text is canonical and follows the rules, or No otherwise.

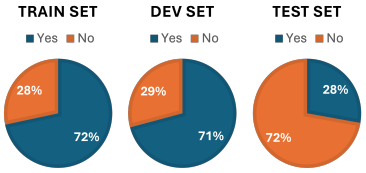

Figure 1: Statistics about the class distribution in the Training, Development and Poetry Test sets. 'Yes' refers to the positive class (i.e. the example is Canonical) and 'No' to the negative one.

For training, we used the VIT Treebank [30], which contains approximately $320,000$ words. Among other information, each sentence is categorized into canonical or not. The dataset was divided into a Training set and a Development set with a 90/10 ratio. The class distribution is shown in Figure 1, where it is evident that the vast majority of the sentences are canonical, reflecting the natural usage patterns of Italian speakers.

We employed the LoRA [31] technique and the Peft package on a single Tesla T4 GPU to train the models for 3 epochs, with a learning rate of 3^{-4} and using a linear scheduler with 10% warmup. The LoRA R parameter was set to 8, α to 16, and all available layers were involved (for more details, refer to the original paper [31]). For computational efficiency, the floating-point precision of the parameters was set to 8 bits, allowing the use of a single GPU.

For the Test set, we used a collection of Italian poetry comprising 51 texts with a total of 303 sentences. For the same reason that people still regard Dante as the greatest Italian poet and students are required to learn his best poems by heart, we have chosen what is regarded as the best Contemporary Italian poetry: a manually curated collection of excerpts from Italian poems from the late 19th and early 20th centuries. In particular, we used poems from the 1975 Nobel Prize Eugenio Montale, with about one hundred excerpts taken from the volume "*Ossi di Seppia*". The class distribution of this test set is shown in Figure 1. Notably, the distribution of Yes (the sentence is canonical) and No (the sentence is non-canonical) is reversed compared to the Training and Development sets, due to poetic license and rhyming constraints. This reversal poses a significant challenge for the models we trained, but it presents an interesting test case. More details about this and a simple Error Analysis are presented in the Appendix B.

In this context, it is important to note that the consideration of structures which, in Chomskyan transformational theory, were once viewed as surface-level realizations of deep canonical structures has not been a deliberate focus of this experiment. The first reason for

Table 1
Classification results on the Test Dataset. FT for each model here refers to the Fine-Tuning procedure, 0s for the 0-shot and 1s for the 1-shot In-context Learning technique.

Model Type	Precision				Recall				F1-Score			
	Yes	No	Macro	Micro	Yes	No	Macro	Micro	Yes	No	Macro	Micro
Yes-Baseline	0,28	0,00	0,14	0,28	**1,00**	0,00	0,50	0,28	0,44	0,00	0,22	0,28
LLaMA1 0s	0,31	0,70	0,51	0,58	0,02	0,90	0,46	0,58	0,03	0,79	0,41	0,58
LLaMA1 1s	0,15	0,70	0,43	0,56	0,27	0,69	0,48	0,56	0,19	0,69	0,44	0,56
LLaMA2 0s	0,28	0,55	0,42	0,48	0,05	0,75	0,40	0,48	0,08	0,63	0,36	0,48
LLaMA2 1s	0,28	0,71	0,50	0,47	0,11	0,65	0,38	0,47	0,26	0,72	0,49	0,59
ExtremITA 0s	0,33	0,68	0,51	0,59	0,11	0,88	0,50	0,59	0,17	0,77	0,47	0,59
ExtremITA 1s	0,27	0,67	0,47	0,49	0,24	0,70	0,47	0,49	0,25	0,68	0,47	0,49
LLaMAntino 0s	0,26	0,74	0,50	0,58	0,12	0,90	0,51	0,58	0,16	0,81	0,49	0,58
LLaMAntino 1s	0,31	0,74	0,53	0,59	0,14	0,85	0,50	0,59	0,19	0,79	0,49	0,59
Camoscio 0s	0,35	0,73	0,54	**0,70**	0,10	**0,93**	0,51	**0,70**	0,15	**0,82**	0,48	**0,70**
Camoscio 1s	0,27	0,72	0,49	0,59	0,26	0,72	0,49	0,59	0,26	0,72	0,49	0,59
BERT FT	0,27	0,70	0,49	0,40	0,67	0,30	0,48	0,40	0,38	0,42	0,49	0,40
Camoscio FT	**0,41**	**0,98**	**0,70**	0,60	**0,98**	0,46	**0,72**	0,60	**0,58**	0,63	**0,60**	0,60

excluding structures like passives, interrogatives, relative clauses, cleft sentences, tough constructions, and others, is their relative scarcity in poetry, though they are more frequent in prose. A second reason, closely tied to the first, is that these common structures do not add an element of surprise, given their frequency in everyday language use. That said, some of these common non-canonical structures can still be found in Italian literary prose, but not all are represented in the examples we studied. On the other hand, focus fronting (also referred to as object preposing, complement preposing, or full argument inversion, depending on the constituent being fronted) is prevalent in the examples included in the experiment. An exemplar list of such structures can be found in Appendix C.

4.1. Results and Discussion

The models used in this paper are those already anticipated in Section 3.2, available from Huggingface, using the prompt described in Section 3.1. The results are available in Table 1. Given the distribution of the sentences of the Training set, we report a simple but informed `Yes-Baseline`. This baseline cannot perform well on the inverted distribution of the Test Set, as it always answers `Yes`. We first used the LLMs anticipated in Section 3.2 in a 0-shot manner and you can notice an overall good ability to detect the non-canonical sentences reaching a 73% of Precision and 93% of Recall for Camoscio, but still struggles to identify the canonical ones. We hoped to heavily boost the performances of the model in the 1-shot scenario[3], but it seemed to decrease in performance. The same trend can be noted for all the other models. As a second comparison, we train an Italian BERT model for 3 epochs which starts showing some awareness of the task and reaching an overall 40% of Micro-F1. Using our Development set we selected only the best LLM to report here for space constraints, which is based on Camoscio [25]. Finally, the Fine-Tuned model reaches the best performance with a very good Precision (98%) for the non-canonical sentences and very good Recall (98%) for the canonical ones, with a final 60% of both Macro and Micro F1.

[3]We experimented with more than 1 example per class, increasing the number of samples up to a 16-shot scenario. Unfortunately, the performance was not increasing but stale around 60% of Micro-F1. We didn't report such results here for space constraints.

4.2. Corpus Analysis

For a better insight into the current measured performance, we studied the role of training material as representative of the adopted test dataset. We analyzed the test dataset used in terms of the average word frequencies, as observed on the ITWaC corpus[4]. This corpus provides pre-computed frequencies for each word: for comparative reasons, we normalized in $[0, 1]$ and measured them for each sentence in terms of the mean frequency, i.e., the sum of the word frequencies over each sentence. By independently averaging frequencies of canonical and non-canonical sentences, we obtained the following figures:

- Canonical Sentences, AVG frequency: 0.38
- Non-Canonical Sentences, AVG frequency: 0.24

Intuitively, a value approaching 1 characterizes highly frequent words in ITWaC: this suggests that they are well-represented in the original LLM. Conversely, values closer to 0 characterize less represented sentences. Notice that only canonical sentences (AVG 0.38) are represented, although in a limited manner, in standard Italian texts. This result sheds light on the specific relationship

[4]https://www.sketchengine.eu/itwac-italian-corpus/

Table 2
Classification results in a 5-fold cross-validation scenario, where the performance for all the splits is merged together.

Model Type	Precision				Recall				F1-Score			
	Yes	No	Macro	Micro	Yes	No	Macro	Micro	Yes	No	Macro	Micro
Yes-Baseline	0,72	0,00	0,36	0,72	**1,00**	0,00	0,50	0,72	0,84	0,00	0,42	0,72
Camoscio FT	**0,93**	**0,83**	**0,88**	**0,90**	**0,93**	**0,84**	**0,88**	**0,90**	**0,93**	**0,83**	**0,88**	**0,90**

between word frequencies and training: LLMs, particularly *Camoscio*, are more "confident" with words they encountered during pre-training or fine-tuning. It is noticeable that almost 50% of our test set words (adjectives, verbs, nouns) do not even occur in the ITWaC and, in fact, they are also absent in any canonical sentence of the training set. Another issue lies in the pre-training data of these LLMs. Since most of the data is in English (over 88%) and non-canonical sentences are extremely rare in English, models like LLaMA or Camoscio have rarely encountered such data, leading to suboptimal performance. Moreover, the length of the sentence could be a factor that may influence the performance of LLMs, specifically in poetry, in the ability to detect canonical or non-canonical sentences.

Therefore, to achieve a more balanced evaluation, we merged the Training, Development, and Testing sets into a single dataset to balance the classes and ensure that the model learns to recognize non-canonical sentences. We then performed an N-Fold Cross-Validation ($N = 5$). Only the trained model was re-evaluated, and the results are presented in Table 2. We maintained the simple and informed `Yes-Baseline` for comparison and recomputed its performance. In this setting, the class distribution aligns again with the Training set. The fine-tuned Camoscio model now shows very good performance in distinguishing canonical sentences, achieving a Macro-F1 of 88% and a Micro-F1 of 90%.

5. Conclusions

In this study, we have shown the potential of Large Language Models, particularly the LLaMA architecture and its Italian adaptations, in distinguishing between canonical and non-canonical sentences in Italian. Our experiments indicate that instruction-tuned models specifically for Italian, such as Camoscio and LLaMAntino, exhibit a strong grasp of Italian syntax and can effectively handle diverse sentence structures. However, the performance for this task is still penalized by the large portion of English data they ingest during pre-training. The findings underscore the importance of tailored language models for specific languages and the benefits of incorporating extensive syntactic variations into training datasets. Future work should focus on expanding the training datasets with more diverse syntactic structures and improving model architectures to better capture the nuances of non-canonical sentences.

Acknowledgments

Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma. We acknowledge financial support from the PNRR MUR project PE0000013-FAIR and support from Project ECS 0000024 Rome Technopole, - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

References

[1] R. Delmonte, Syntax and semantics of italian poetry in the first half of the 20th century, 2018. URL: https://arxiv.org/abs/1802.03712. arXiv:1802.03712.

[2] R. Delmonte, Cognitive Models of Poetry Reading, Springer International Publishing, Cham, 2021, pp. 1–39. URL: https://doi.org/10.1007/978-3-030-44982-7_19-4. doi:10.1007/978-3-030-44982-7_19-4.

[3] R. Delmonte, Recursion and Ambiguity: A Linguistic and Computational Perspective, 2015, pp. 257–284. doi:10.1007/978-3-319-08043-7_15.

[4] J. Bresnan, The Mental Representation of Grammatical Relations, The MIT Press, Cambridge, 1982.

[5] J. Bresnan, Lexical-Functional Syntax, Blackwell Publishing, Oxford, 2001.

[6] G. Ward, B. Birner, Information Structure and Non-canonical Syntax, 2008, pp. 152 – 174. doi:10.1002/9780470756959.ch7.

[7] R. Delmonte, N. Busetto, Measuring similarity by linguistic features rather than frequency, in: H. Bunt (Ed.), Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 42–52. URL: https://aclanthology.org/2022.isa-1.6.

[8] C. Bosco, V. Lombardo, D. Vassallo, L. Lesmo, Building a treebank for Italian: a data-driven annotation schema, in: M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stain-

hauer (Eds.), Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), European Language Resources Association (ELRA), Athens, Greece, 2000. URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/220.pdf.

[9] C. Bosco, A. Mazzei, V. Lombardo, G. Attardi, A. Corazza, A. Lavelli, L. Lesmo, G. Satta, M. Simi, Comparing Italian parsers on a common treebank: the EVALITA experience, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/528_paper.pdf.

[10] S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte, Building the Italian Syntactic-Semantic Treebank, Springer Netherlands, Dordrecht, 2003, pp. 189–210. URL: https://doi.org/10.1007/978-94-010-0201-1_11. doi:10.1007/978-94-010-0201-1_11.

[11] T. Chung, M. Post, D. Gildea, Factors affecting the accuracy of Korean parsing, in: D. Seddah, S. Koebler, R. Tsarfaty (Eds.), Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, Association for Computational Linguistics, Los Angeles, CA, USA, 2010, pp. 49–57. URL: https://aclanthology.org/W10-1406.

[12] C. Bosco, A. Mazzei, A. Lavelli, Looking back to the evalita constituency parsing task: 2007-2011, in: B. Magnini, F. Cutugno, M. Falcone, E. Pianta (Eds.), Evaluation of Natural Language and Speech Tools for Italian, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 46–57.

[13] T. Paccosi, A. Palmero Aprosio, S. Tonelli, It is markit that is new: An italian treebank of marked constructions, in: CLiC-it 2021 - Italian Conference on Computational Linguistics, 2022.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the NAACL 2019, 2019, pp. 4171–4186.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019).

[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020).

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019).

[23] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] C. D. Hromei, D. Croce, R. Basili, U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models, IJCoL 10 (2024). URL: http://journals.openedition.org/ijcol/1352.

[25] A. Santilli, E. Rodolà, Camoscio: an Italian Instruction-tuned LLaMA, 2023. URL: https://arxiv.org/abs/2307.16456. arXiv:2307.16456.

[26] P. Basile, E. Musacchio, M. Polignano, L. Siciliani,

G. Fiameni, G. Semeraro, LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language, 2023. URL: https://arxiv.org/abs/2312.09993. arXiv:2312.09993.

[27] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, 2024. URL: https://arxiv.org/abs/2301.00234. arXiv:2301.00234.

[28] D. Croce, A. Moschitti, R. Basili, Structured lexical similarity via convolution kernels on dependency trees, in: R. Barzilay, M. Johnson (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 1034–1046. URL: https://aclanthology.org/D11-1096.

[29] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[30] R. Delmonte, A. Bristot, S. Tonelli, VIT – Venice Italian Treebank: Syntactic and Quantitative Features, in: Proc. Sixth International Workshop on Treebanks and Linguistic Theories, volume 1, Nealt Proc. Series, 2007, pp. 43–54. URL: https://catalog.elra.info/en-us/repository/browse/ELRA-W0324/.

[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021).

A. Limitations

In assessing the data distribution disparities between languages in the pre-training phase of the LLaMA family models, we provide an illustrative breakdown in Table 3, where English accounts for nearly 90% of the data, while Italian is present in less than 1%.

Among the limitations of the proposed model, the computational costs associated with training a model like LLaMA are undoubtedly significant, requiring hundreds of hours on a GPU. We have implemented methods to streamline this process, but the computational expenditure for training on a 16GB GPU remains high. This becomes even more pronounced considering the model's sentence processing time, which is slightly less than half a second per sentence. Given the required computational power to run the model, this duration is relatively long.

Table 3
Data distribution.

Code	Language	Percentage
en	English	**89,70%**
unk	unknown	8,38%
de	German	0,17%
fr	French	0,16%
sv	Swedish	0,15%
zh	Chinese	0,13%
ru	Russian	0,13%
es	Spanish	0,13%
nl	Dutch	0,12%
it	Italian	0,11%
ja	Japanese	0,10%
pl	Polish	0,09%
pt	Portuguese	0,09%
vi	Vietnamese	0,08%
uk	Ukrainian	0,07%
ko	Korean	0,06%
ca	Catalan	0,04%
sr	Serbian	0,04%
cs	Czech	0,03%
fi	Finnish	0,03%
hu	Hungarian	0,03%
id	Indonesian	0,03%
no	Norwegian	0,03%
ro	Romanian	0,03%
bg	Bulgarian	0,02%
da	Danish	0,02%
hr	Croatian	0,01%
sl	Slovenian	0,01%

Regarding the model's application, since it heavily relies on an LLM, it might be susceptible to hallucination — generating non-existent sentences or fragments. However, during inference (few-shot or training), it seems to always answer in the request format, very rarely (especially in 0-shot) adding some explanation for its decision after a Yes or No.

Additional experiments might be necessary to ensure that pollution effects don't unduly influence the evaluation process: the VIT dataset might have been encountered during the pre-training phase. Although this might have occurred, certainly the model did not have the opportunity to observe sentences from the poetry domain associated with the canonical or non-canonical label.

B. Error Analysis

In this section, we present a simple Error Analysis with two different cases: *i)* a sentence from the Development set, which should reflect the distribution of the training data for the models introduced in Section 3.2; a sentence from the poetry domain that is radically different from the training data. We will then report the answer for each model specifying the modality (in-context learning or training) and eventually the number of shots used for inference.

As a first example, consider "*Difficile tenersi in quel cammino*"[5], which is non-canonical as the main verb "è" is missing. The models answered as follows:

- LLaMA1 0s: canonical
- LLaMA1 1s: canonical
- LLaMA2 0s: canonical
- LLaMA2 1s: canonical
- ExtremITA 0s: canonical
- ExtremITA 1s: non-canonical
- LLaMAntino 0s: canonical
- LLaMAntino 1s: non-canonical
- Camoscio 0s: non-canonical
- Camoscio 1s: non-canonical
- BERT FT: non-canonical
- Camoscio FT: non-canonical

This example is interesting because all the Italian adapted models in some way (1-shot or Fine-Tuned) answered correctly, thus recognizing that the sentence was missing the main verb, given the initial prompt. Notice that only Camoscio answered correctly both in 0-shot and 1-shot

As a second and more difficult example, consider the sentence "*Zacinto mio che te specchi nell'onde del greco mar da cui vergine nacque Venere*"[6], taken from the poetry test set. This example is very hard to comprehend as some words are very rare in spoken/written Italian (*nell'onde*), the usage of the uncommon *te* to express that the city is actively mirroring in the sea, and the reversed order of the last words. In this case, all the models answered that the sentence is non-canonical, recognizing the strange structure of the sentence, except for BERT FT which classified this sentence as canonical.

C. Typical Non-Canonical Structures

In this section, we report a list of typical non-canonical structures as an example of the complexity the models are dealing with.

1. Inversion of the complete argument, where the complement is fronted, and the subject follows the verb.
2. Subject inversion, positioning the subject after the main verb.
3. Fronting of the object, moving the object to the beginning of the sentence before the subject.
4. Extraction of the object from an infinitival clause, placing it at the beginning of the sentence.
5. Preposing of a prepositional adjunct from a participial clause, moving the prepositional complement of a past participle to a position before the verb.
6. Leftward extraction of the lexical verb, where the untensed, non-finite main verb precedes the auxiliary or modal verb.
7. Right dislocation of the subject, placing the subject after the complements of the sentence.
8. Fronting of both the subject and the object, positioning them before the main verb, with the subject preceding the object.
9. Fronting of a prepositional specification, often introduced by "of", extracting it from the noun phrase and positioning it at the front.
10. Right dislocation of the clitic, where a clitic pronoun attached to the main verb corefers to an object noun phrase positioned later in the sentence.
11. Right dislocation of the object, placing the object after indirect objects, adjuncts, or an inverted subject.
12. Insertion of parentheticals or adjuncts between the subject and the main verb.
13. Rightward extraction of the adjective from the noun phrase, positioning it after any noun adjuncts.
14. Right stranding of a prepositional specification, such as "of", leaving it at the end of the sentence, separate from the noun phrase.
15. Rightward extraction of the lexical verb, positioning the untensed, non-finite main verb after the complements of the sentence.
16. Right stranding of the predicate's head noun, leaving it after two adjuncts.

[5] In English: "*(It's) Hard to keep in that path.*"
[6] In English: "*My Zacinto that you mirror in the waves of the Greek sea where virgin was born Venus from*"