Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns?

James A. Michaelov
Department of Cognitive Science
University of California, San Diego
j1michae@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California, San Diego
bkbergen@ucsd.edu

Abstract

Some languages allow arguments to be omitted in certain contexts. Yet human language comprehenders reliably infer the intended referents of these *zero* pronouns, in part because they construct expectations about which referents are more likely. We ask whether Neural Language Models also extract the same expectations. We test whether 12 contemporary language models display expectations that reflect human behavior when exposed to sentences with zero pronouns from five behavioral experiments conducted in Italian by Carminati (2005). We find that three models—XGLM 2.9B, 4.5B, and 7.5B—capture the human behavior from all the experiments, with others successfully modeling some of the results. This result suggests that human expectations about coreference can be derived from exposure to language, and also indicates features of language models that allow them to better reflect human behavior.

1 Introduction

In Italian, like other *pro-drop* ('pronoun-dropping') languages, verbal arguments that would usually be expressed by pronouns in languages such as English can be omitted under certain circumstances. For example, consider the sentence in (1) from Carminati (2005).

(1) *Quando Maria ha chiamato Mario, era contenta*.
'When Maria called Mario, [she] was happy.'

In this sentence, the referent of the 'dropped' pronoun—generally referred to as a *zero* or *null* pronoun—can be inferred from the fact that the adjective *contenta* is feminine; thus, Maria is the most likely subject of the second clause. Resolving the referents of anaphoric zero pronouns like this is a long-standing, important, and active area of research in natural language understanding (Jurafsky and Martin, 2021; for examples, see Zhao and Ng, 2007; Taira et al., 2008; Imamura et al., 2009; Watanabe et al., 2010; Kong and Zhou, 2010; Poesio et al., 2010; Chen and Ng, 2013; Yoshino et al., 2013; Iida et al., 2016; Aloraini and Poesio, 2020; Song et al., 2020a; Ueda et al., 2020; Konno et al., 2020, 2021; Yang et al., 2020; Kim et al., 2021; Umakoshi et al., 2021; Chen et al., 2021; Yang et al., 2022).

It has been argued that aiming for human-likeness in natural language processing systems is vital if we want our natural language understanding systems to behave not only as humans do, but also as human users expect them to (see, e.g., Keller, 2010; Ettinger, 2020; Eisape et al., 2020). This is particularly true for zero anaphora resolution, and pronoun resolution more generally. As an illustration of the prominence of reference resolution, one pronoun resolution task, the Winograd Schema Challenge (Levesque et al., 2012; based on work by Winograd, 1972), has been referred to as 'an alternative to the Turing Test' (Levesque et al., 2012, p. 552).

So, how do humans resolve coreference? The evidence suggests that we use a range of cues—for example, agreement information as in (1), but also factors such as world knowledge and common sense (Winograd, 1972; Hobbs, 1979; Kehler et al., 2007; Kehler and Rohde, 2013; Sakaguchi et al., 2019). In addition, pronoun resolution is shaped by our expectations about the next entity that is likely to be mentioned and what argument it should take (Kehler et al., 2007; Kehler and Rohde, 2013; Nieuwland and Van Berkum, 2006). For example, crosslinguistic work has found a bias towards expecting a subject pronoun to refer to an antecedent subject (for discussion, see Carminati, 2005). This has been demonstrated experimentally with sentences such as those in (2).

(2) *John seized the comic from Bill. He____*

When presenting experimental participants with

Proceedings of the 29th International Conference on Computational Linguistics, pages 1–14
October 12–17, 2022.

sentences such as (2) where the content following the pronoun has been removed, Stevenson et al. (1994) found that the vast majority of people expect *he* to refer to *John* rather than *Bill*. The effect of expectations such as these are so powerful that we may often not even realize that a sentence is grammatically ambiguous at all in most situations (Nieuwland and Van Berkum, 2006).

The same principles apply in zero anaphora resolution. Carminati (2005), for example, tests human expectations by investigating how long it takes for experimental participants to read stimuli with certain linguistic features, based on the well-established finding that contextually expected words are read faster than unexpected words, demonstrating an increased processing difficulty when these expectations are violated (see, e.g. Forster, 1981; Levy, 2008; Luke and Christianson, 2016; Brothers and Kuperberg, 2021). Carminati (2005) finds that the main clauses of sentences such as (1)—that is, the part of the sentence containing the zero subject pronoun, i.e., *era contenta* ('[she] was happy')—are read faster when the zero pronoun co-refers with a subject antecedent, as in (1), than when it co-refers with the antecedent object, as in (3).

(3) *Quando Maria ha chiamato Mario, era contento*.
'When Maria called Mario, [he] was happy.'

The question, then, if we want human-like zero anaphora resolution, is how to test whether a given zero anaphora resolution system is able to reflect these human expectations. In the present study, we propose a method to do just that.

The vast majority of recent pronoun resolution systems base their approach around using the representations learned by contemporary transformer language models—for example, in the zero pronoun anaphora resolution literature alone, researchers have used pretrained transformers such as monolingual (Song et al., 2020b; Ueda et al., 2020; Konno et al., 2020, 2021; Kim et al., 2021; Chen et al., 2021; Umakoshi et al., 2021) and multilingual (Aloraini and Poesio, 2020; Kim et al., 2021) BERT models (Devlin et al., 2019), as well as XLM-R (Conneau et al., 2020; for an example see Yang et al., 2022).

For these systems, there is a clear way to test for human-like-ness. We can directly probe the extent to which the representations learned by the language models take into account the factors that lead to coreference expectations in humans by testing how similar the predictions of language models are to those of human comprehenders—if they exhibit the same pattern of predictive behavior as humans in a given context, this demonstrates that they are sensitive to the same factors as humans in this context. We do this by comparing the reading times reported by Carminati (2005) to the surprisals of 12 contemporary transformer language models (Devlin et al., 2019; Conneau and Lample, 2019; De Mattei et al., 2020; Schweter, 2020; Conneau et al., 2020; de Vries and Nissim, 2021; Lin et al., 2021) for the same stimuli.

A key question is whether it is possible for language models to learn any of these human-like expectations at all, given that they can only rely on the statistics of language. For this reason, the results of the present study should be of interest both from a natural language understanding perspective, as discussed above, and also from a psycholinguistics perspective.

From the natural language understanding perspective, the present study presents an approach for 'pre-evaluating' a language model's suitability as a basis for a zero anaphora resolution system. Specifically, if a language model can model a specific effect in human language processing—that is, if an experimental manipulation that elicits a significant difference in reading time also results in a significant difference in that language model's surprisal in the same direction—this demonstrates that it is able to take into account the relevant factors that underlie human comprehender expectation. For example, if a language model can successfully model the subject antecedent preference, this suggests that it has learned that all else being equal, subject antecedents are more likely to be the coreferents of zero subject pronouns, and thus, crucially, that this pattern is in some way represented in the contextual embeddings that can be used as the representations underlying a zero anaphora resolution system.

From the psycholinguistics perspective, this study explores the extent to which it is possible that specific patterns in zero anaphora coreference expectations can be learned on the basis of the statistics of language alone. There is substantial work demonstrating that some expectations are highly correlated with language statistics, and thus may be at least partly derived from them (Levy, 2008; Monsalve et al., 2012; Smith and Levy, 2013;

Frank et al., 2015; Michaelov and Bergen, 2020; Szewczyk and Federmeier, 2022). However, other work has suggested that coreference expectations are instead (or in addition) at least partly based on semantic knowledge, world experience, and conceptual salience (Hobbs, 1979; Harley and Ritter, 2002; Carminati, 2005; Kehler et al., 2007; Kehler and Rohde, 2013). Nonetheless, since the predictions of language models are derived from language statistics alone, if even one language model can successfully model a given effect (after adjusting for multiple comparisons), this provides in-principle evidence that the effect can be successfully learned using distributional information alone.

2 General Method

The experiments reported by Carminati (2005) were self-paced reading experiments. Participants were native speakers of Italian asked to read Italian sentences on a computer. Stimuli were similar to those discussed in the previous section, with a subordinate clause (e.g. *Quando Maria ha chiamato Mario*; 'When Maria called Mario') first presented, followed by the main clause (e.g. either *era contenta* '[she] was happy' or *era contento* '[he] was happy'). The time taken by participants to read the main clause—which includes the word that disambiguates the null subject pronoun—was recorded.

To measure the language model's expectations, we used surprisal (negative log-probability) based on a large body of evidence that language model surprisal generally correlates well with reading time (see, e.g. Levy, 2008; Monsalve et al., 2012; Smith and Levy, 2013; Goodkind and Bicknell, 2018) and other metrics of processing difficulty that are thought to correlate with human expectations such as the neural N400 response (Frank et al., 2015; Aurnhammer and Frank, 2019; Michaelov and Bergen, 2020; Merkx and Frank, 2021).

To model each effect, we compared whether specific linguistic features of the stimuli that elicited a significant difference in human reading times also led to a significant difference in language model surprisal. For example, we investigate whether, like reading time, surprisal is significantly lower when the referent of a zero subject pronoun is a subject antecedent compared to an object antecedent, among other patterns in reading time reported by Carminati (2005). The language models were all presented with the same stimuli as the human participants, which are provided by Carminati (2005) in an appendix to the original paper.

To match reading time, surprisal was calculated over the whole of the main clause in each stimulus item. This was done by calculating the sum of the surprisals of the main clauses' constituent words, which is equivalent to taking the negative logarithm of the product of their probabilities.

We ran the stimuli through 12 transformer language models—5 monolingual and 7 multilingual. Two of the monolingual models were autoregressive transformer networks: GePpeTto (De Mattei et al., 2020) and the small English GPT-2 retrained on Italian (de Vries and Nissim, 2021). The three remaining monolingual models were masked language models: UmBERTo (Parisi et al., 2021) trained on the Italian subcorpus of OSCAR, and the Base and XXL versions of the Italian BERT models (Schweter, 2020). The multilingual models also included autoregressive and masked language models. The autoregressive models were three different sizes of XGLM (Lin et al., 2021): the 2.9B, 4.5B, and 7.5B parameter models. The masked language models were XLM-100 (Conneau and Lample, 2019), and the Base and Large versions of XLM-R (Conneau et al., 2020).

The aim in using this range of models was to test whether there are any model types or characteristics made them better suited to capturing human behavior—for instance, whether the models were autoregressive or masked, or monolingual or multilingual. Previous systems designed to resolve zero pronoun anaphora of the kind described here appear to be predominantly based on masked language models; however, autoregressive models such as GPT-2 have been successfully used in similar systems (Maqbool et al., 2022).

We are also interested in whether monolingual or multilingual models are better suited to the task of zero pronoun anaphora resolution—while cross-lingual transfer may help with some phenomena (Guarasci et al., 2022), there is also evidence that it can cause harm to model performance in others (Wang et al., 2020). There is currently mixed evidence with respect to zero pronoun anaphora resolution—Kim et al. (2021), for example, find that a monolingual Korean BERT-based model performs better than the standard multilingual BERT model; while Yang et al. (2022) finds that their model, based on XLM-R, is better than a model based on a Chinese-only BERT (Song et al., 2020b). We include both multilingual BERT and XLM-R

in our analyses, in addition to the Base Italian BERT model, which has previously been evaluated in terms of its capacity to learn non-anaphoric null subject and agreement phenomena in Italian (Guarasci et al., 2021).

To test whether each model successfully modeled each effect, we constructed linear mixed-effects models predicting model surprisal with experimental manipulation as a main effect and a random intercept of sentence frame, where sentence frame refers to a set of stimuli that differ only by experimental condition (e.g., the previously discussed *Quando Maria ha chiamato Mario, era contenta* and *Quando Maria ha chiamato Mario, era contento* are two stimuli with the same sentence frame).

For three of the five analyses—the two analyses in Section 3.1 where the coreferent is distinguished by gender, and the analysis in Section 3.2—we tested whether the relevant experimental manipulation was a significant predictor of language model surprisal by constructing a null regression with only the random intercept of sentence frame and running a likelihood ratio test investigating whether adding the experimental manipulation improved model fit.

The remaining two analyses correspond to two different tests utilized by Carminati (2005) to analyze the results of a single experiment (Experiment 4 of the original paper). Crucially, Carminati (2005) tests whether there is an interaction between coreferent argument (whether it is the antecedent subject or object) and coreferent person (whether the coreferent is in the first or second person or in the third person), but also whether there is a main effect of each of these. To test whether there is an interaction (in Section 3.3), we construct a linear mixed-effects model with and without the interaction, and run a likelihood ratio test comparing the two. In addition to the interaction, Carminati (2005) finds a main effect of coreferent argument but not of person. Thus, we also test for the main effect of coreferent argument, which we report in Section 3.1. Because we want to investigate whether the main effect of coreferent argument explains a significant amount of the variance in surprisal while also accounting for the effect of a possible interaction, instead of using a likelihood ratio test, we opt for a Type III ANOVA with Satterthwaite's method for estimating degrees of freedom (Kuznetsova et al., 2017).

The details of the results of the statistical analyses that were run by Carminati (2005) are provided in the original paper. The full results of the statistical analyses that we ran are provided in Appendix A. The results of both sets of statistical analyses are summarized in Figure 1.

All language models were run in *Python* (Van Rossum and Drake, 2009), using the *PyTorch* (Paszke et al., 2019) implementation of each model, as provided by the *transformers* package (Wolf et al., 2020). Statistical analysis and data manipulation were carried out in *R* (R Core Team, 2020) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), *ggsignif* (Ahlmann-Eltze and Patil, 2021), and *cowplot* (Wilke, 2020) packages. The stimuli, code used to run the models, and code used to run the statistical analyses are provided on Github[1]. Note that all p-values reported in this analysis have been corrected for multiple comparisons (Benjamini and Hochberg, 1995; R Core Team, 2020).

3 Manipulation-level results and discussion

In this section, we compare the performance of the 12 language models tested with human behavior on five of the experimental manipulations carried out by Carminati (2005). Note that two additional studies from that paper focus on a different question—the effects of distractor referents on processing time. Although at least one model was able to capture each of these human results, they are not included here because they address a different set of phenomena.

3.1 Subject vs. object antecedent referent

Carminati (2005) investigates the subject antecedent preference discussed in Section 1 in three experiments. In Experiments 1 and 2 of the original paper, both antecedents are names associated with different genders, as illustrated by the example from Experiment 1 shown in (4).

(4) (a) *Quando Lucia ha telefonato a Marco, era appena tornata da Londra.*
'When Lucia has telephoned Marco, [she] had just come back from London.'

(b) *Quando Lucia ha telefonato a Marco, era appena tornato da Londra.*
'When Lucia has telephoned Marco, [he] had just come back from London.'

[1] https://github.com/jmichaelov/italian-zero-anaphora-prediction

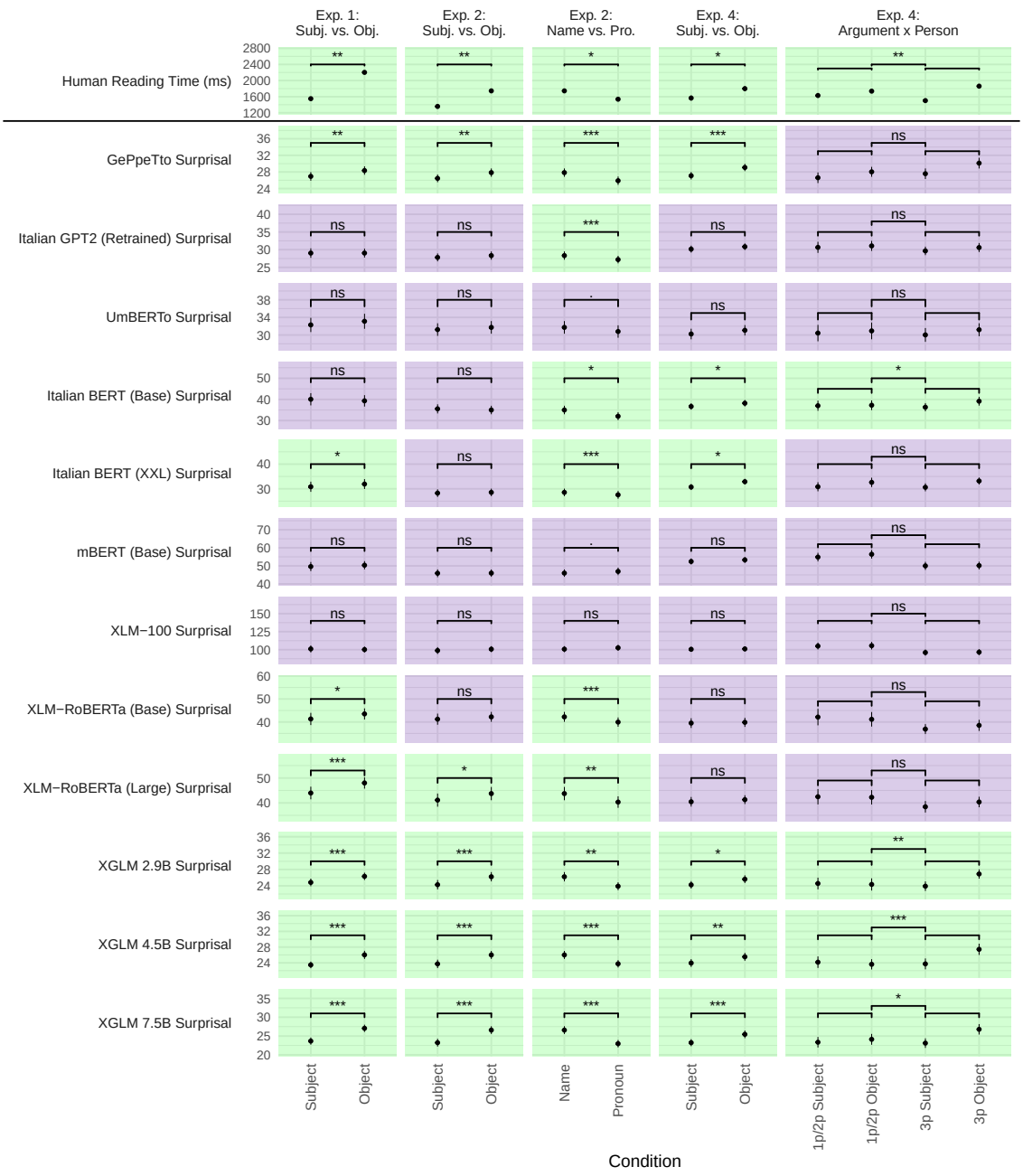

Figure 1: Mean reading time and surprisal of each model elicited by main clauses for each experimental condition in each experiment. All significant differences are shown: following convention, '***' indicates $p < 0.001$, '**' indicates $p < 0.01$, '*' indicates $p < 0.5$, '.' indicates marginal significance where $p < 0.1$, and 'ns' indicates $p \geq 0.1$. For easier comparison across models and experiments, comparisons with statistically significant results are colored green; non-significant results are colored purple. Note that the relevant p-values have been corrected for multiple comparisons using the method of Benjamini and Hochberg (1995); for test statistics and degrees of freedom, see Appendix A. Details of the statistical tests for reading time are provided by Carminati (2005). For language model surprisal, error bars indicate standard error; no metric of error is provided by Carminati (2005).

Because *tornato/tornata* ('come back') agrees with the gender of the zero subject pronoun, its referent can be resolved to be the subject antecedent (*Lucia*) in (4a) and the object antecedent (*Marco*) in (4b). Carminati (2005) found, as expected, that main clauses where the zero subject pronoun co-referred with the subject antecedent (like (4a)) were read faster than those where they had an object antecedent coreferent (like (4b)), suggesting an expectation for a subject antecedent coreferent.

In Experiment 4 of the original paper, grammatical person was manipulated rather than grammati-

cal gender, as illustrated by the example in (5).

(5) (a) *Quando ho litigato con Maria, ero molto prepotente.*
'When [I] quarrelled with Maria, [I] was very pushy.'

(b) *Quando ho litigato con Maria, era molto prepotente.*
'When [I] quarrelled with Maria, [she] was very pushy.'

Similarly, because *ero/era* ('was') either agrees with the first person or third person, the zero subject pronoun can be resolved as co-referring with the speaker (in (5a)) or with Maria (in (5b)). As in the aforementioned other experiments, Carminati (2005) finds that speakers read sentences like (5a) faster than sentences like (5b), again demonstrating a preference for subject antecedent coreferents over object antecedent coreferents.

Looking at the results of the models, we can see that only GePpeTto and the XGLMs successfully model this effect in all three experiments. This appears to suggest that autoregressive models may be better at learning that the subject antecedent is the more likely referent; however, it should be noted that in each of the individual studies, at least one masked language model also successfully modeled the effect. Nonetheless, the robustness of similarity between these autoregressive models' predictions and human expectations may be partly explained by the evidence suggesting that autoregressive models are more sensitive to word order than masked language models, to the extent that they are able to encode positional information even without explicit positional encodings (Haviv et al., 2022); conversely, masked language models appear to be relatively insensitive to word order (Sinha et al., 2021; Gupta et al., 2021). Given that the dominant pattern in Italian is Subject-Verb-Object (see Guarasci et al., 2022) and the subject was always first in the subordinate clause, it is therefore unsurprising that autoregressive models would be better able to predict that the first entity mentioned (the subject) is more likely as the subject of the zero pronoun than the second entity mentioned (the object).

3.2 Name vs. pronoun antecedent referent

In addition to investigating the differences in how humans process zero anaphora in sentences with subject and object antecendent coreferents, Carminati (2005) also investigated how the form in which antecedents are presented impacts processing. As a further part of Experiment 2 of the original paper, Carminati (2005) investigates how processing is impacted when the object coreferent is presented as a name or a pronoun, an example of which is provided in (6).

(6) (a) *Quando Maria cerca Roberto, diventa ansioso.*
'When Maria looks for Roberto, [he] becomes anxious.'

(b) *Quando Maria lo cerca, diventa ansioso.*
'When Maria looks for him, [he] becomes anxious.'

In both sentences, it is the object antecedent that is the referent of the zero pronoun in the main clause, violating the subject antecedent preference. Carminati (2005) finds that main clauses with zero pronouns referring to antecedent objects are easier to process (read faster) when this antecedent object is a pronoun.

The results for the language models, shown in Figure 1, suggest that this is a relatively easy pattern for language models to learn—9 of the 12 models show a significant effect and the remaining 3 show a marginal effect in the correct direction. Thus it is clear that this general rule—that an entity referred to by an antecedent pronoun is more likely to be the referent of a zero pronoun—is possible to learn based on the statistics of language. The fact that this effect relies on the form of the antecedents rather than word order could explain why there is no difference between autoregressive and masked language models in this case.

3.3 Antecedent argument by grammatical person interaction

In addition to investigating the subject antecedent effect, in Experiment 4 of the original paper, Carminati (2005) investigates how this effect interacts with the the grammatical person of antecedents (i.e., first, second, or third-person). In general, previous work suggests that first and second-person antecedents are more likely to be referents of reduced or zero pronouns (Ariel, 1991; Siewierska, 1999, 2003; Carminati, 2005), but as has been discussed, subject antecedents are also more likely to be their referents. Thus, Carminati (2005) compares the effect of the person of the coreferent antecedent when it is in both subject and object position, as exemplified in (7).

(7) (a) *Quando ho/hai litigato con Maria, ero/eri molto prepotente.*
'When [I/you] quarrelled with Maria, [I/you] was/were very pushy.'

(b) *Quando Maria ha litigato con me/te, ero/eri molto prepotente.*
'When Maria quarrelled with me/you, [I/you] was/were very pushy.'

(c) *Quando Maria ha litigato con me/te, era molto prepotente.*
'When Maria quarrelled with me/you, [she] was very pushy.'

(d) *Quando ho/hai litigato con Maria, era molto prepotente.*
'When [I/you] quarrelled with Maria, [she] was very pushy.'

While Carminati (2005) does not find a main effect of grammatical person, the results show an interaction between person and antecedent referent argument status (i.e. whether it is a subject or object). Specifically, the difference in reading time between subject and object antecedent referents is reduced when the antecedent coreferent is in the first or second person. In other words, the subject antecedent effect is weaker with first and second person coreferents. This, Carminati (2005) argues, shows that the bias towards a first or second-person coreferent modulates the bias against an object coreferent—in other words, humans still expect a first or second-person coreferent even if it is an object antecedent.

Four of the models—Italian BERT Base and the XGLMs—manage to model this interaction. While this suggests the the effect—which is complicated as it relies on correctly weighting the effects of argument status and person—is difficult to learn based on the statistics of language, it nevertheless demonstrates that it is indeed possible.

4 General Discussion

4.1 Implications for human language processing

We can now return to the two questions that motivated this work. First, we look at whether the reading time effects in humans can be explained on the basis of the statistics of language.

As seen in Figure 1, each experimental result was successfully modeled by at least four language models, after correcting for multiple comparisons. This shows that it is possible to learn cues based on the statistics of language that result in human-like expectations about the referents of zero subject pronouns in Italian. The fact that the XGLM transformers were consistently able to model all the effects demonstrates that the patterns underlying the results of the experiments can all be learned by the same system—and therefore, in principle, it should also be possible for a neurocognitive system implementing lexical prediction in humans (for accounts of such a system and what it might learn, see, e.g., Kutas et al., 2011; Lewis and Bastiaansen, 2015; Lupyan and Clark, 2015; Frank et al., 2015; Bornkessel-Schlesewsky and Schlesewsky, 2019; Aurnhammer and Frank, 2019; Michaelov and Bergen, 2020; Kuperberg et al., 2020; Merkx and Frank, 2021; Brothers and Kuperberg, 2021). Thus, the present study provides evidence that the expectations that humans form about possible referents in anaphora may be derived from language statistics, at least in part.

4.2 Implications for work on language models

Model	Experiments modeled
GePpeTto	4/5
It. GPT2 (Retrained)	1/5
UmBERTo	0/5
It. BERT (Base)	3/5
It. BERT (XXL)	3/5
mBERT	0/5
XLM-100	0/5
XLM-R (Base)	2/5
XLM-R (Large)	3/5
XGLM 2.9B	5/5
XGLM 4.5B	5/5
XGLM 7.5B	5/5

Table 1: Number of experiments successfully modeled by each language model.

The number of experiments successfully modeled by each language model is shown in Table 1, revealing that the XGLM models performs best overall, successfully modeling the results of all 5 experiments investigated. After the XGLMs, GePpeTto models the most experiments (4/5), followed by XLM-R Large and the Italian BERTs (3/5). The remaining transformers only successfully model 2 or fewer of the experiments.

At this level of analysis, some patterns begin to emerge. First, the best models are the XGLM transformers and GePpeTto. This suggests that autore-

gressive models may in fact be best able to model the effects. As discussed in Section 3.1, this may be due to their comparatively high sensitivity to word order. One issue that confounds this interpretation is that the XGLM models are also larger and trained more data on than the other models. However, the fact that GePpeTto was trained on 13GB of text, while the other monolingual models (which were all masked language models) were trained on the same amount or more data and performed worse, suggests that, at the very least, monolingual autoregressive models may more efficiently learn biases in zero anaphora processing than monolingual masked language models. Whether or not autoregressive models continue to out-perform masked language models as the training set increases in size is a question for further research. Overall, then, we see that in our sample of models, autoregressive monolingual and multilingual models are more human-like in their expectations of zero subject pronoun referents than their masked language model counterparts.

Another question that we can address with the present results is that of the effect of multilinguality on the human-likeness of the models' expectations. First, while GePpeTto and Retrained Italian GPT-2 are trained on the same Italian corpus, the former greatly out-performs the latter. This suggests that training a model on one language and then re-training it on another does not necessarily improve the representations that a model learns—in fact, in this case, it interferes with the model's ability to make predictions in a human-like fashion. On the other hand, XLM-R Large is trained on data from 100 languages successfully models human processing at least as well as any monolingual model but GePpeTto—including Italian BERT XXL, which is trained on 80GB of Italian text compared to XLM-R's 30GB. Thus, it may be the case that with more training data, and with a larger number of languages (including more closely-related languages—XLM-R is also trained on other Romance languages), there is some cross-linguistic transfer that can aid in predicting the referent of a null subject pronoun in a human-like manner (see Guarasci et al., 2022, for a recent similar finding). Finally, the XGLMs—autoregressive multilingual models—are the best performing models overall. Thus, the results of this study seem to suggest that with enough overall data, and when multilingual language models are trained on more languages, cross-linguistic transfer can improve their human-likeness in terms of their predictions. A question for future work is to investigate under what circumstances multilinguality hurts or harms the human-likeness of language model predictions—for example, based on how related the languages the model is trained on are to each other, or how widespread the phenomenon under investigation is. For example, the subject antecedent preference is also present in English with overt pronoun anaphora (Smyth, 1994; Chambers and Smyth, 1998; Kehler et al., 2007; Kehler and Rohde, 2013).

Finally, as discussed in Section 2, Yang et al. (2022) show that a zero pronoun anaphora resolution system based on XLM-R performs better than one based on multilingual BERT (Song et al., 2020b). Concurrently, in the present study, we see that either XLM-R model is better able to model zero anaphora processing effects than multilingual BERT. While there are other factors at play, this result is consistent with our prediction that better modeling of human expectations may lead to better performance when using the models' representations for zero pronoun anaphora resolution, based on the idea that the representations learned by the model better allow it to make human-like predictions, and thus are more useful for systems aiming to resolve zero anaphora in a human-like way. In the present study, XGLM models perform better than the other models, and thus, based on this, we suggest that XGLM transformers may be better models upon which to base future zero pronoun anaphora resolution system than other current publicly available pretrained models.

5 Conclusion

We present the first study investigating whether language models make the same predictions as humans when processing zero pronoun anaphora. For each the 5 effects we investigate, we find that there are at least four models that successfully do so; and three models, XGLM 2.9B, 4.5B, and 7.5B, successfully do so in all 5. This suggests that human processing of zero pronoun anaphora may at least partly rely on our statistical knowledge of language. Furthermore, this approach provides a useful way to investigate how human-like the referent predictions of language models are, which is vital if we are to use their representations for zero anaphora resolution systems.

Acknowledgements

We would like to thank Maria Nella Carminati for making her stimuli available, and the anonymous reviewers for their helpful comments. We would also like to thank the other members of the Language and Cognition Lab at UCSD for their valuable discussion and the University of California, and the San Diego Social Sciences Computing Facility Team for technical assistance. This work was partially supported by a 2021-2022 Center for Academic Research and Training in Anthropogeny Annette Merle-Smith Fellowship awarded to James A. Michaelov, and the RTX A5000 used for this research was donated by the NVIDIA Corporation.

References

Constantin Ahlmann-Eltze and Indrajeet Patil. 2021. Ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'.

Abdulrahman Aloraini and Massimo Poesio. 2020. Cross-lingual Zero Pronoun Resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 90–98, Marseille, France. European Language Resources Association.

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5):443–463.

Christoph Aurnhammer and Stefan L. Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. 2019. Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, 10.

Trevor Brothers and Gina R. Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Maria Nella Carminati. 2005. Processing reflexes of the Feature Hierarchy (Person > Number > Gender) and implications for linguistic theory. *Lingua*, 115(3):259–285.

Craig G. Chambers and Ron Smyth. 1998. Structural Parallelism and Discourse Coherence: A Test of Centering Theory. *Journal of Memory and Language*, 39(4):593–608.

Chen Chen and Vincent Ng. 2013. Chinese Zero Pronoun Resolution: Some Recent Advances. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1360–1365, Seattle, Washington, USA. Association for Computational Linguistics.

Shisong Chen, Binbin Gu, Jianfeng Qu, Zhixu Li, An Liu, Lei Zhao, and Zhigang Chen. 2021. Tackling Zero Pronoun Resolution and Non-Zero Coreference Resolution Jointly. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 518–527, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. GePpeTto Carves Italian into a Language Model: Italian Conference on Computational Linguistics 2020. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Bologna, Italy. CEUR-WS.org.

Wietse de Vries and Malvina Nissim. 2021. As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze Distillation: Improving Neural Language Models with Human Next-Word Prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Allyson Ettinger. 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.

K. I. Forster. 1981. Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):465–495.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Assessing BERT's ability to learn Italian syntax: A study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Computer Speech & Language*, 71:101261.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & Family Eat Word Salad: Experiments with Text Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12946–12954.

Heidi Harley and Elizabeth Ritter. 2002. Person and Number in Pronouns: A Feature-Geometric Analysis. *Language*, 78(3):482–526.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer Language Models without Positional Encodings Still Learn Positional Information.

Jerry R. Hobbs. 1979. Coherence and Coreference*. *Cognitive Science*, 3(1):67–90.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-Sentential Subject Zero Anaphora Resolution using Multi-Column Convolutional Neural Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1244–1254, Austin, Texas. Association for Computational Linguistics.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 85–88, Suntec, Singapore. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2021. *Speech and Language Processing*, third edition. [Online Draft].

A. Kehler, L. Kertz, H. Rohde, and J. L. Elman. 2007. Coherence and Coreference Revisited. *Journal of Semantics*, 25(1):1–44.

Andrew Kehler and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Frank Keller. 2010. Cognitively Plausible Models of Human Language Processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden. Association for Computational Linguistics.

Youngtae Kim, Dongyul Ra, and Soojong Lim. 2021. Zero-anaphora resolution in Korean based on deep language representation model: BERT. *ETRI Journal*, 43(2):299–312.

Fang Kong and Guodong Zhou. 2010. A Tree Kernel-Based Unified Framework for Chinese Zero Anaphora Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA. Association for Computational Linguistics.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo Zero Pronoun Resolution Improves Zero Anaphora Resolution. *arXiv:2104.07425 [cs]*.

Ryuto Konno, Yuichiroh Matsubayashi, Shun Kiyono, Hiroki Ouchi, Ryo Takahashi, and Kentaro Inui. 2020. An Empirical Study of Contextual Data Augmentation for Japanese Zero Anaphora Resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4956–4968, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gina R. Kuperberg, Trevor Brothers, and Edward W. Wlotko. 2020. A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, 32(1):12–35.

Marta Kutas, Katherine A. DeLong, and Nathaniel J. Smith. 2011. A look around at what lies ahead: Prediction and predictability in language processing. In Moshe Bar, editor, *Predictions in the Brain: Using Our Past to Generate a Future*, pages 190–207. Oxford University Press, New York, NY, US.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1–26.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Ashley G. Lewis and Marcel Bastiaansen. 2015. A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, 68:155–168.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

Gary Lupyan and Andy Clark. 2015. Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, 24(4):279–284.

M.H. Maqbool, Luxun Xu, A.B. Siddique, Niloofar Montazeri, Vagelis Hristidis, and Hassan Foroosh. 2022. Zero-label Anaphora Resolution for Off-Script User Queries in Goal-Oriented Dialog Systems. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 217–224.

Danny Merkx and Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A. Michaelov and Benjamin K. Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2006. Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, 1118(1):155–167.

Loreto Parisi, Simone Francia, and Paolo Magnani:. 2021. UmBERTo Commoncrawl Cased.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Massimo Poesio, Olga Uryupina, and Yannick Versley. 2010. Creating a Coreference Resolution System for Italian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. RStudio, PBC., Boston, MA.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv:1907.10641 [cs]*.

Stefan Schweter. 2020. Italian BERT and ELECTRA models. Zenodo.

Anna Siewierska. 1999. Reduced pronominals and argument prominence. In *Proceedings of the LFG99 Conference*, page 14, University of Manchester. CSLI Publications.

Anna Siewierska. 2003. Reduced pronominals and argument prominence. In Miriam Butt and Tracy Holloway King, editors, *Nominals: Inside and Out*, Studies in Constraint-Based Lexicalism. CSLI Publications, Stanford, Calif.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Ron Smyth. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23(3):197–229.

Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. 2020a. ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5429–5434, Online. Association for Computational Linguistics.

Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. 2020b. ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5429–5434, Online. Association for Computational Linguistics.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Jakub M. Szewczyk and Kara D. Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese Predicate Argument Structure Analysis using Decision Lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 523–532, Honolulu, Hawaii. Association for Computational Linguistics.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based Cohesion Analysis of Japanese Texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese Zero Anaphora Resolution Can Benefit from Parallel Texts Through Neural Transfer Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. 2010. A Structured Model for Joint Learning of Argument Roles and Predicate Senses. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 98–102, Uppsala, Sweden. Association for Computational Linguistics.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Claus O. Wilke. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingxuan Yang, Si Li, Sheng Gao, and Jun Guo. 2022. CorefDPR: A Joint Model for Coreference Resolution and Dropped Pronoun Recovery in Chinese Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:571–581.

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Ji-Rong Wen, and Nianwen Xue. 2020. Transformer-GCRF: Recovering Chinese Dropped Pronouns with General Conditional Random Fields. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 137–147, Online. Association for Computational Linguistics.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2013. Predicate Argument Structure Analysis using Partially Annotated Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 957–961, Nagoya, Japan. Asian Federation of Natural Language Processing.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages

541–550, Prague, Czech Republic. Association for Computational Linguistics.

A Full results of statistical analyses

Experiment 1: Subject vs. object antecedent referent

Model	Chisq(df=1)	Corrected *p*
GePeTto	**12.3**	**0.002**
GPT-2 Italian	<0.1	0.993
UmBERTo	3.5	0.109
It BERT Base	0.4	0.614
It BERT XXL	**6.3**	**0.025**
mBERT	1.4	0.376
XLM-100	0.2	0.758
XLM-R Base	**5.3**	**0.041**
XLM-R Large	**19.2**	**<0.001**
XGLM 2.9B	**13.9**	**<0.001**
XGLM 4.5B	**19.1**	**<0.001**
XGLM 7.5B	**32.9**	**<0.001**

Table 2: Results of the likelihood ratio tests in Experiment 1. Models for which there is a significant effect of the manipulation tested are shown in bold.

Experiment 2: Subject vs. object antecedent referent

Model	Chisq(df=1)	Corrected *p*
GePeTto	**8.8**	**0.008**
GPT-2 Italian	0.9	0.482
UmBERTo	0.7	0.514
It BERT Base	0.9	0.482
It BERT XXL	0.6	0.570
mBERT	<0.1	0.956
XLM-100	0.9	0.482
XLM-R Base	1	0.468
XLM-R Large	**7.5**	**0.015**
XGLM 2.9B	**17.5**	**<0.001**
XGLM 4.5B	**23.6**	**<0.001**
XGLM 7.5B	**26.5**	**<0.001**

Table 3: Results of the likelihood ratio tests for all models in Experiment 2.1. Models for which there is a significant effect of the manipulation tested are shown in bold.

Experiment 2: Name vs. pronoun object antecedent referent

Model	Chisq(df=1)	Corrected *p*
GePeTto	**29.8**	**<0.001**
GPT-2 Italian	**17.3**	**<0.001**
UmBERTo	4.9	0.050
It BERT Base	**7.6**	**0.015**
It BERT XXL	**17.5**	**<0.001**
mBERT	4.3	0.068
XLM-100	2.7	0.168
XLM-R Base	**14.7**	**<0.001**
XLM-R Large	**11.7**	**0.002**
XGLM 2.9B	**12.4**	**0.002**
XGLM 4.5B	**16.4**	**<0.001**
XGLM 7.5B	**26.6**	**<0.001**

Table 4: Results of the likelihood ratio tests for all models. Models for which there is a significant effect of the manipulation tested are shown in bold.

Experiment 4: Subject vs. object antecedent referent

Model	F(1,60)	Corrected *p*
GePeTto	**34.6**	**<0.001**
GPT-2 Italian	1.5	0.359
UmBERTo	1.1	0.434
It BERT Base	**7.4**	**0.019**
It BERT XXL	**9.1**	**0.010**
mBERT	0.7	0.529
XLM-100	<0.1	0.815
XLM-R Base	<0.1	0.830
XLM-R Large	0.5	0.575
XGLM 2.9B	**7.1**	**0.022**
XGLM 4.5B	**12**	**0.003**
XGLM 7.5B	**18.4**	**<0.001**

Table 5: Results of the ANOVAs for all models. Models for which there is a significant effect of the manipulation tested are shown in bold.

Experiment 4: Argument x Person Interaction

Model	Chisq(df=1)	Corrected *p*
GePeTto	2.7	0.168
GPT-2 Italian	0.2	0.709
UmBERTo	0.2	0.739
It BERT Base	**5**	**0.048**
It BERT XXL	0.3	0.681
mBERT	0.4	0.607
XLM-100	<0.1	0.993
XLM-R Base	1.1	0.434
XLM-R Large	0.8	0.482
XGLM 2.9B	**8.9**	**0.008**
XGLM 4.5B	**18.5**	**<0.001**
XGLM 7.5B	**7.4**	**0.015**

Table 6: Results of the likelihood ratio tests for all models. Models for which there is a significant effect of the manipulation tested are shown in bold.