Probing for Dutch Relative Pronoun Choice

Bouma, Gosse

Published in:
Computational Linguistics in the Netherlands Journal

Publication date:
2021

Computational Linguistics in the Netherlands Journal 11 (2021) 59-70 Submitted 10/2021; Published 12/2021

Probing for Dutch Relative Pronoun Choice

Gosse Bouma G.BOUMA@RUG.NL

University of Groningen

Abstract

We propose a linguistically motivated version of the relative pronoun probing task for Dutch (where a model has to predict whether a masked token is either *die* or *dat*), collect realistic data for it using a parsed corpus, and probe the performance of four context-sensitive BERT-based neural language models. Whereas the original task, which simply masked all occurrences of the words *die* and *dat*, was relatively easy, the linguistically motivated task turns out to be much harder. Models differ considerably in their performance, but a monolingual model trained on a heterogeneous corpus appears to be most robust.

1. Introduction

Large neural language models such as BERT (Devlin et al. 2019) are trained among others on a masked language prediction task, where the input consists of sentences where some portion of the input words have been masked, and the model has to predict which of several candidate words are the most likely to appear in the masked positions. Statistical, ngram-based, language models as well as early neural models such as word2vec (Mikolov et al. 2013) only take a very limited context into account when making predictions about a following or masked token. Transformer-based neural language models such as BERT in theory are sensitive to the full (left and right) context of a masked position when making predictions. One question that arises is whether this allows these neural models to make accurate predictions in cases where the correct answer depends on making grammatical distinctions that depend on a larger context (Linzen and Baroni 2021). For instance, in a subject-verb agreement task (Linzen et al. 2016), one might probe the model for its ability to distinguish between singular and plural agreement in cases such as (1), where the model has to be sensitive to the fact that it is the number of the head noun of the subject phrase, and not that of the immediately preceding noun, that determines the form of the verb.

(1) The length of the forewings (is/*are) ...

The overview of probing research in Linzen and Baroni (2021) shows that results from such experiments are mixed. On the one hand, models display an accuracy that suggests that they are sensitive to part of speech and syntactic structure, while on the other hand there are also results that show that performance quickly decreases in syntactically complex cases. For instance, in sentences where the subject is modified by a relative clause (2), or where the first noun in the sentence is not heading the subject (3), models are less good at predicting the correct form of the verb. Apart from subject-verb agreement, inflectional morphology (Haley 2020), negative polarity (Warstadt et al. 2019), subject-auxiliary-inversion (McCoy et al. 2020), and sensitivity to long-distance dependencies involving 'gaps' and island constraints (Wilcox et al. 2018, Wilcox et al. 2021) have also been used to probe models. Sahin et al. (2019) provide a multilingual suite of probing tasks.

(2) The movie that the author likes, (is/*are) good

(3) The scientist thinks that parts of the river valley have/*has

Delobelle et al. (2020) test RobBERT, a BERT-based neural language model for Dutch, and a number of competing language models, on a masked language prediction task where the model has to predict for a given sentence whether the masked token is either *die* or *dat*. These two words can be used, among others, as relative pronouns, where *dat* can only occur with an antecedent that is a singular neuter (*het*) noun (4-a), and *die* can be used with either a singular non-neuter antecedent (*de*) (4-b) or a plural antecedent.

(4) a. het portret van een vrouw dat bij hem op de ezel stond
the portrait of a woman that with him at the easel stood
the portrait of a woman that stood at his easel
b. de directeur van een bedrijf die in een Duitse auto rijdt
the director of a company that in a German car drives
the director of a company that drives a German car

The task is inspired by a similar experiment from Allein et al. (2020). Allein *et al.* use data from the Europarl corpus[1] and Sonar[2] to train a neural classifier for *die/dat*-prediction, using an LSTM that is initialized with word embeddings obtained from a word2vec model. The trained classifier obtains accuracies of 83.2% (Europarl) and 84.5% (Sonar). Delobelle et al. collect data from the Dutch section of the Europarl corpus, where they use all sentences containing *die* or *dat* as test cases (288K sentences). They report accuracies of 90.2 (mBERT, a multilingual BERT model), 94.9 (Bertje, an alternative Dutch language model, (de Vries et al. 2019)) and 98.7% (RobBERT). It should be noted, though, that in many cases, there is no real ambiguity or need to pay attention to longer contexts. In (5-a), for instance, *dat* is used as subordinating conjunction in a position where *die* could never occur. In (5-b) *die* is used as relative pronoun with *gegevens* as antecedent, but there are no preceding nouns that could function as distractor. In (5-c), *die* is used as deictic pronoun introducing an NP where the noun with which the pronoun agrees is adjacent to the pronoun. Even if there is intervening material in such cases, it usually is not a noun that could act as distractor. This task is therefore not optimal for 'probing', as it is unclear to what extent information from the full context of the masked position is required to make the correct prediction.

(5) a. U begrijpt die/**dat** we te maken hebben met een bijzondere prestatie
You understand that we to do have with a special achievement
You understand that we are dealing with a special achievement
b. De gegevens **die**/dat u invult
The data that you enter
c. De toestand van **die**/dat wegen is slecht
The condition of these roads is bad

We propose to turn the *die/dat* prediction task into a proper probing task by focusing on cases where a relative pronoun occurs in the masked position (thus ignoring prediction of the relatively easy subordinate conjunction and deictic pronoun cases) and where there is at least one distractor. We explain how we collected relevant examples from a parsed corpus, and test two monolingual and two multilingual neural language models for their ability to make the right predictions and thus establish to what extent such models are sensitive to longer contexts.

2. Relative Clause Attachment

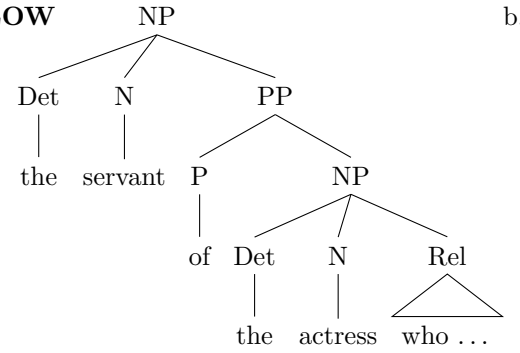
Relative clause attachment has been studied extensively from a psycholinguistic perspective (starting with Cuetos and Mitchell (1988)), also for Dutch (Desmet et al. 2006). The canonical example is:

1. https://www.statmt.org/europarl/
2. https://taalmaterialen.ivdnt.org/download/tstc-sonar-corpus/

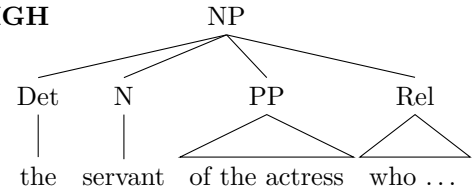
(6) Someone shot the servant of the actress who was on the balcony

Here, the relative clause could be attached either to the lower noun *actress* as in (7-a) or the higher noun *servant* as in (7-b). It has been claimed that preference for high or low attachment is not universal, with speakers of some languages having a preference for low attachment, where other languages, such as Dutch, prefer high attachment (Brysbaert and Mitchell 1996). Desmet et al. point out that corpus data shows that in Dutch, low attachment is the most frequent configuration as well, and high attachment only is more frequent in cases where the highest noun is animate. The preference of test subjects is in line with this observation if one takes these factors into account during construction of the test items.

(7) a. **LOW**



b. **HIGH**



Whereas psycholinguistic studies have used a sentence completion task to probe the preference for test subjects for either high or low attachment of the relative clause, we change the task into a masked word prediction task suitable for testing a language model by only masking the relative pronoun.[3] As the language model is trained on raw text and never sees any syntactic structures, we exclusively concentrate on those cases where the two preceding nouns in the complex noun phrase have opposite gender (*neuter* or *non-neuter*, respectively), thus triggering either *die* or *dat*, which we take as signs of high or low attachment in this particular context.

(8) a. Het is het proza$_{neu}$ van een vrouw$_{nonn}$ [MASK] een hoge prijs betaalde
It is the prose of a woman [MASK] a high price paid
It is the prose of a woman that paid a high price

b. Melk is het enige product$_{neu}$ van de koe$_{nonn}$ [MASK] aan de paniek is ontsnapt
Milk is the only product of the cow [MASK] from the panic is escaped
Milk is the only product of the cow that has escaped from the panic

Example (8-a) illustrates low attachment, where the relative clause is modifying the lower noun *vrouw*, and the correct pronoun is *die*$_{nonn}$. (8-b) illustrates high attachment, where the relative clause modifies *product*, and the masked pronoun therefore is *dat*$_{neu}$.

It should be noted that our probing experiment serves a very different purpose than the reading comprehension experiments. The psycholinguistic research has concentrated on answering the question whether there is an inherent preference for low or high attachment by speakers of a given language, and has used specifically constructed test items, in which context ideally does not make either one of the attachment options much more plausible than the other. Properly controlling for confounding factors remains an issue, however (Hemforth et al. 2015). We use models trained on a corpus of naturally occurring data and also test on corpus data. In such data, most relevant cases, even when there are two potential antecedents, will be such that one antecedent is clearly more plausible than the other. Statistics indicate that 'low attachment' in general is much more

3. See Ettinger (2020) for experiments where psycholinguistic datasets are directly used to test the sensitivity of a language model to semantic roles and context.

frequent than 'high attachment' in actual corpus data, and that, for Dutch, the relative pronoun *die* occurs more often than *dat*. Thus, a language model trained on such data is likely to develop a preference for low attachment (i.e. selecting the closest preceding noun to predict the form of the relative pronoun), and for predicting *die* over *dat*, but a model that is sensitive to context should still be able to predict high attachment (i.e. implictly selecting the leftmost noun as antecedent) or *dat* as pronoun in cases where the alternative would lead to semantic incoherence. Note that we adopt the terminology from phrase structure syntax in describing the two cases as 'high' and 'low' attachment respectively, but that the models are not trained on syntactically annotated data and incorporate syntactic knowledge only to a certain extent (but see Manning et al. (2020)). Instead, a successful model will learn which nouns typically go with which kinds of relative clauses, and use that information to decide on the correct pronoun.

3. Data Collection

To collect realistic examples, we searched a newspaper corpus, containing articles from *Algemeen Dagblad* and *NRC Handelsblad* 1994-1995. This corpus has not been used during training of any of the language models we use in the experiment, and thus there cannot be memorization effects (Carlini et al. 2020). The corpus was automatically parsed with Alpino (van Noord 2006). Our corpus query searched for sentences containing a complex NP with the structure of either (7-a) or (7-b) and with the additional constraints that the two relevant nouns had to be singular and of opposite gender and that the embedded PP had to be headed by *van* (*of*). The latter constraint was added to reduce the number of false hits due to parsing errors, as *van*-PPs almost always modify a noun and occur rarely as modifiers of verbs. We did not impose any other constraints. In particular, any number of other modifying elements (adjectives, other PPs) can be present, and the relative clause may be extraposed.

The Alpino parser uses a statistical model obtained from a corpus to disambiguate sentences, and, as low attachment is more frequent than high attachment, tends to do better on low attachment cases than on high attachment cases. This is reflected in the extracted sentences for the high attachment case, as these contain a substantial number of false hits. Therefore, we manually selected 1951 true positives from the 2534 hits returned by the query. A brief inspection of the low-attachment cases showed that these hardly contain false hits. We therefore did not filter these. As low attachment is also much more frequent than high attachment and we wanted to use balanced data, we included 2000 instances of low attachment. Note that while false hits (and duplicate sentences) were removed, sentences that actually are ambiguous after masking the pronoun, such as (9), were kept.

(9) ... een dagboek van de nieuwe topman [MASK] een tip oplicht van de geheimen van de
a diary of the new head [MASK] a tip reveals of the secrets of the
onderhandeling .
negotiation
a diary of the new head that reveals the secrets of the negotiation

Here, the choice for *dat* amounts to choosing *dagboek* as antecedent (in line with the source) whereas *die* amounts to choosing *topman* as antecedent, which is a semantically plausible interpretation as well. The datasets and code used in this paper are available at github.[4]

4. Language Models

We tested on two multilingual and two monolingual context-sensitive language models (Table 1). Whereas statistical (ngram) language models are typically trained to make a prediction about the next token in a sentence, neural language models are trained using a masked-language model (MLM)

4. https://github.com/gossebouma/die_vs_dat

Model	Corpus	Training objective	Reference
RobBERT	Dutch section of OSCAR (web, 39GB)	MLM	(Delobelle et al. 2020)
Bertje	Dutch Wikipedia, Sonar, Novels (12GB)	MLM	(de Vries et al. 2019)
mBERT	Wikipedia for 104 languages	MLM and NSP	Github[a]
XLM-R	Common Crawl (web, multilingual, 2.5TB)	MLM	(Conneau et al. 2020)

Table 1: Neural-language models used in the probe.

a. https://github.com/google-research/bert/blob/master/multilingual.md

objective, where some words in the text are masked and the model has access to both left and right context to make a prediction. All models of the BERT-variety (Devlin et al. 2019) use a complex neural architecture that includes an attention mechanism that allows the model to learn which words in the input are most important for making a prediction about the masked position. The question thus arises whether this enables the model to learn that, say, the head noun inside a complex subject NP is the relevant word for predicting the correctly inflected form of the finite verb, or whether the model instead learns a simpler but structure insensitive rule (such as attending to the closest preceding noun) that would allow it to make the right prediction in most cases without paying attention to structure. Similarly, for the *die/dat* probing task as we have designed it, the model must decide which of two preceding nouns is more likely as antecedent, where the following context (containing the relative clause) may contain crucial information for making the correct decision. Some models also use a next-sentence prediction (NSP) task for training, where the model is trained to predict whether two concatenated sentences occurred in the given order in the corpus or not. While this can be relevant for downstream tasks such as question answering, it is probably less relevant for learning to attend to linguistic structure. A language model can be trained on a monolingual corpus, but recently it has been shown that multilingual models, trained on a concatenation of monolingual corpora (with upsampling of the data for low-resource languages), can outperform monolingual models, especially for low-resource languages (Pires et al. 2019). We used the pretrained models present on huggingface.co[5] and tested the models on a MLM task where the relative pronoun was replaced by [MASK] and the model assigns a probability to both *die* and *dat* as fillers. The accuracy of the model is the percentage of test items where the model assigns the highest probability to the pronoun that matches with the gold standard.

In a second experiment, the relative pronoun prediction task was framed as a classification task, where the model has to predict for a given sentence (in which the relative pronoun is masked), whether it is an instance of a sentence containing '*die*' or '*dat*'. For this classification experiment, the original models were fine-tuned on a set of over 130.000 sentences containing a relative pronoun, but without any further constraints on the context in which the pronoun occurs.

While the MLM experiment measures to what extent models are able to predict 'out of the box' the form of the relative pronoun, the classification experiment measures to what extent models are able to learn this task if given a suitable amount of training examples.

5. Experiments and Discussion

In order to do well on the relative pronoun prediction task, a model must take both left and right context into account. The left context provides two nouns, while the right context contains the relative clause. In order to decide on the correct pronoun, the model must work out whether the relative clause is more likely to be attached to the higher or lower noun in the left context. The results in Table 2 show that the 'high attachment' cases are much harder than 'low attachment' for

5. https://huggingface.co/GroNLP/bert-base-dutch-cased, https://huggingface.co/pdelobelle/robbert-v2-dutch-base, https://huggingface.co/xlm-roberta-base, https://huggingface.co/bert-base-multilingual-cased

	High attachment					Low attachment				
	N	Bertje	RobBERT	mBERT	XLM-R	N	Bertje	RobBERT	mBERT	XLM-R
dat	888	**0.789**	0.537	0.510	0.590	1000	**0.931**	0.782	0.769	0.852
die	1163	0.740	0.650	0.711	**0.783**	1000	0.937	0.836	0.885	**0.957**
total	1951	**0.761**	0.601	0.624	0.700	2000	**0.934**	0.809	0.827	0.905

Table 2: Probing results (accuracy) for the MLM experiment, predicting the correct relative pronoun in 'high' and 'low' attachment configurations without any fine-tuning

	High attachment					Low attachment				
	N	Bertje	RobBERT	mBERT	XLM-R	N	Bertje	RobBERT	mBERT	XLM-R
dat	888	**0.714**	0.672*	0.616*	0.702*	1000	0.953*	0.944*	0.950*	**0.961***
die	1163	**0.704**	0.682*	0.594	0.666	1000	**0.966***	0.960*	0.954*	0.957
total	1951	**0.708**	0.678*	0.604	0.682	2000	**0.960***	0.952*	0.952*	0.959*

Table 3: Probing results (accuracy) for the classification experiment, predicting the correct relative pronoun in 'high' and 'low' attachment configurations after fine-tuning. Results that improve over the scores in Table 2 are marked with an asterisk.

all models. This is not unexpected, as low attachment is more frequent in the corpus from which the test sentences are extracted and probably in the training corpora as well, and high attachment requires the model to attend to a noun that is relatively far from the masked position with an intervening noun that has the opposite gender. Bertje is the only model that does equally well on the *die* and *dat* cases, where the other models all have a tendency to prefer *die* over *dat*. Apart from Bertje, the multilingual model XLM-R suffers least from the difference between *die* and *dat* and therefore does better than the other two models.

As an alternative to probing language models 'out of the box' on a task that requires to predict the correct form of a masked input token, one may also test to what extent a model can be fine-tuned on this particular task. In that second setting, the model parameters are fine-tuned by having it predict the correct label ('*die*' or '*dat*') given an input consisting of a sentence where, as in the masked language modeling set-up, the relative pronoun is masked.

As training material, a set of 130K sentences containing the relative pronoun '*die*' or '*dat*' was collected from one section (NH1994) of the CLEF corpus. All sentences that also occur in the test data were removed. Note that this training set contains occurrences of relative pronouns in general, and that only a small portion of the training material consists of sentences that are structurally similar to the test data.[6] While, after training[7], prediction accuracies on the validation data (similar to the training data) is very high for all models (over 98% accuracy), performance on the more challenging test data is as shown in table 3.

Comparing the results for the various models without and with fine-tuning shows that for low attachment all models benefit from fine-tuning, and that the models that performed weakest without fine-tuning benefit most, so that the differences between them are much smaller (less than 1%) after fine-tuning. For the high attachment cases, which are rare in the training data, the effect is mixed. Although Bertje still obtains the highest accuracy, it performs substantially worse than without fine-tuning, while RobBERT improves substantially. For the two multilingual models, the effect of fine-tuning on the high-attachment cases is slightly negative. Thus, it seems that the effect of

6. In particular, as we removed duplicates, only structurally cases consisting of a complex NP with two nouns and a relative, where the lower noun is governed by a preposition other than '*van*' are included.
7. for all models we trained for 3 epochs, with learning rate $2e^{-05}$, and Adam optimizer

fine tuning is that it improves scores for low attachment examples in particular, while this is not consistently the case for the high-attachment cases. This is most likely an effect of the training data, which samples the distribution for relative pronouns in general, and which most likely contains only a small portion of relevant high-attachment cases.

5.1 Error Analysis

We manually inspected the sentences where Bertje (without fine-tuning) made a wrong prediction. We assume that the distribution of errors for the other models will be similar. We distinguished between cases where the system made a wrong prediction that leads to an incoherent sentence (*Error*), cases where the alternative pronoun and antecedent is not excluded by the context (*Ambiguous*), and cases where something else may lead to the wrong pronoun being predicted (*Other*). The latter includes cases where a neuter noun denotes a human being (where both *dat* and *die* are acceptable) and cases where there is an additional distractor.

Ambiguous cases are examples such as (10), where both the first round or the tournament can be held according to modified rules (*dat* appears in the test sentence, but the model predicts *die*).

(10) de eerste ronde van het bekertoernooi, dat/die in gewijzigde opzet zal worden
the first round of the cup-tournament that in modified modus will be
afgewerkt
held
the first round of the tournament, that will be held in a modified form

Other cases are among other examples with a human antecedent, as in (11). Neuter nouns with a human antecedent are an exception to grammatical pronoun agreement, in that they can be referred to by both *die* and *dat*. In this case the test sentence has *dat* where the model predicts *die*.

(11) de dood van haar buurjongetje Esajas dat/die van het dak gevallen is
the death of her boy-nextdoor Esajas that of the roof fallen is
the death of the boy nextdoor Esajas, who is fallen of the roof

Intervening distractors can also cause mismatches, such as in (12), where the neuter noun *aeroshell* is an apposition to the first noun *hitteschild* (non-neuter), making the model prefer *die* over *dat*. In this case, both options lead to a grammatical and semantically equivalent sentence.

(12) het hitteschild van 2 meter diameter, de aeroshell, dat/die de instrumenten moet
the heat-shield of 2 meter diameter, the aeroshell, that the instruments must
beschermen
protect
a heat-sheald of 2 meters wide, the aeroshell, that must protect the instruments

Table 4 shows that for the low attachment cases, the number of ambiguous cases is almost the same as the number of 'real' errors (i.e. errors leading to an unacceptable sentence). For the high-attachment cases, the number of real errors is higher, probably because performance on this configuration is also worse than on low-attachment cases. The number of errors due to some other reason in the low attachment cases is slightly higher than in the high attachment errors, probably a consequence of the fact that the low-attachment cases were not manually filtered.

As errors on the ambiguous cases, and some fraction of the other cases, do not lead to semantically or syntactically incoherent sentences, one might argue that the reported accuracies are actually a bit pessimistic, as some of the cases that are counted as an error are actually linguistically acceptable. Another indication for the fact that true ambiguity and presence of neuter nouns denoting a human influences results, is the fact that humans, when asked to do the same task as the model, also do not perform at 100% accuracy. In a small experiment, we asked three subjects to predict the masked

		N	Error	Ambig	Other
LOW	die	61	25	25	11
	dat	72	29	26	17
HIGH	die	100	76	22	2
	dat	100	55	38	7

Table 4: Error analysis for Bertje of low attachment errors and two 100 sentence samples of high attachment errors

pronoun for a balanced sample of 200 sentences. All annotators noticed that it was a hard task that required careful reading of the, often long and complex, sentences and feared that they might not have done well. Performances varied between and 75% and 95%. While this result should be seen as very preliminary, it does suggest that it is a hard task for humans as well.

5.2 Left and Right Context Regression Testing

One might wonder to what extent left or right context alone is a good predictor for the task. For instance, assuming that low attachment is the default, the model might use the observation that high attachment is much more likely for some N_1 *van* N_2 combinations. If the model learns what the most likely pronoun is for such combinations, it might do well on the task without taking right context into account. We suspect that this is true for instance for somewhat idiomatic expressions such as *een politiek symbool van de eerste orde (a political sign of great significance), de meest vraatzuchtige zeug van het westelijke halfrond (the most hungry pig of the western hemisphere)*, and *een terrein van grote zorg (an area of great concern)* and cases where the second noun denotes an amount or temporal or monetary unit such as *de periode van mijn leven (the period of my life), een vrouw van een jaar of zeventig (a woman approximately seventy years old), het probleem van deze week (the problem of this week), een kind van deze eeuw (a child of this century).* But although such cases do occur in the data, they do not seem to be especially frequent.

It might also be that the relative clause contains indicators that either *die* or *dat* is far more likely. If that is the case, a model could make accurate predictions by only taking right context into account. In the majority of examples, the relative pronoun has the function of subject in the relative clauses (13-a). Object relatives (13-b) are far less frequent. In subject relatives, the pronoun (and its antecedent) have to be possible as subjects of the verb involved.

(13) a. de basis van ieder team dat iets wil bereiken
the base of every team that something wants achieve
the foundation of every team that wants to achieve something
b. de taal van het land dat ik bezoek
the language of the country that I visit
the language of the country that I visit

Of the 2273 verb forms that occur as head of a relative clause in the newspaper corpus we used for obtaining the test sentences, only 66 occur with *dat* more often than *die*. They include forms such as *aangeboden (offered), gepresenteerd (presented), voorziet (anticipates), opgesteld (written), verschijnt (appears), gesloten (closed).* They are mostly passive participles for which the grammatical subject is usually inanimate and thus the probability of a neuter noun as antecedent of the relative pronoun goes up.

		Actual	Synthetic data	
			matching	non-matching
LOW	*dat*	0.931	0.802	0.792
	die	0.937	0.851	0.831
	overall	0.934	0.827	0.811
HIGH	*dat*	0.789	0.595	0.551
	die	0.740	0.622	0.560
	overall	0.761	0.608	0.555

Table 5: Results for Bertje on actual and synthetic data.

While these observations suggest that there might be some cases where right or left context alone is a good predictor of the correct pronoun, it does not directly answer the question to what extent both right and left context are required to perform well on the task. As an alternative, we tested to what extent performance degrades if left and right context do not form a semantically coherent sentence. In particular, we created synthetic data in which a left context, including the two nouns and the relative pronoun (which we consider as 'gold' answer for this test), is combined with a right context from another, randomly selected, test item. There are two cases, one where the right context was from a test item that had the same relative pronoun as the left context, and one where this is not the case. I.e. in 'matching' cases of low attachment with pronoun *dat*, we created a new example by taking the left context (containing the two nouns) and combining it with a random right context (containing the relative clause), but still starting with *dat*. The result is often semantically incoherent, as in (14). In the 'non-matching' cases, we combined a left context ending with *dat* with a right context originally starting with *die*, as in (15). The predictions for Bertje (without fine-tuning) on these data-sets are shown in table 5.

(14) Aan de voorgevel van het gebouw [MASK] zich van meet af aan sterk heeft
At the front of the building [MASK] itself from beginning of at strong has
gemaakt voor de Tiger
made for the Tiger
At the front of the building [which] argued for the Tiger from the beginning

(15) Het indexcijfer van de industriële produktie [MASK] vanmorgen tientallen Palestijnen
The index of the industrial production [MASK] morning tens Palestinians
in de moskee heeft vermoord
in the mosque has killed
The index of the industrial production [which] killed tens of Palestinians in the mosque this morning

For synthetic data, the model is considerably less capable of predicting the correct pronoun (again, assuming that the left context provides the 'correct' answer) than for the original corpus material. This means that both left and right context are important for making the correct prediction, and that data in which left and right contexts do not form a semantically coherent input, are harder for the model than original data. Although 'non-matching' synthetic examples are a bit harder than 'matching' examples, the difference between the two is small compared to the difference with actual data. This suggests that the relative clauses alone do not provide a strong signal for predicting either *dat* or *die*, and that the model is confused as soon as a semantically incoherent sentence is encountered, independent of the original pronoun in the relative clauses.

6. Conclusion

Transformer-based neural language models are able to perform remarkably well on masked language prediction tasks. To see to what extent such models are sensitive to linguistic structure, we reformulated the *die/dat* prediction task for Dutch. Inspired by previous work in psycholinguistics, we collected a corpus of test sentences where a relative clause could in theory be attached to one of two preceding nouns, and the model has to be able to select the correct antecedent noun in order to make the correct prediction. There are significant differences in performance between the four BERT models that we probed. The Bertje model did best in the masked language modeling probe, but with a monolingual model doing best and worst on the task, it cannot be concluded that monolingual models are always to be preferred over multilingual models (or vice versa). Results after fine-tuning are mixed, with improved accuracy scores (compared to the MLM task) for low-attachment cases both with lower scores for the more challenging high-attachment cases. While Bertje still performs best after fine-tuning, differences between the models are small and the RobBERT model in particular seems to benefit from fine-tuning.

An important factor that influences performance might be the corpus used for training the language models. The sentences used in the probe are all relatively complex, due to the fact that we select sentences with a complex NP containing a PP and a relative clause. Such sentences might be more frequent in corpora that include newspaper text than in corpora consisting of text from Wikipedia or the web only. In the newspaper corpus we used, we estimate that around 80% of the relevant cases are instances of low attachment. It might well be that in other corpora the distribution is even more skewed, making it increasingly hard for the model to learn to make the right predictions for high attachment cases. The non-neuter relative pronoun *die* is somewhat more frequent than *dat* (approx. 60% and 40%, respectively) in our data. It seems that this cannot explain the tendency for some models to prefer *die*, although this preference might be stronger in more recent corpora, especially if they also contain informal text (Audring 2013, Bouma 2017).

The fact that we can test for high or low attachment by simply masking the pronoun is due to the fact that Dutch relative pronouns agree with the antecedent. In light of the discussion from the psycholinguistics literature on (speakers of) languages having a preference for either high or low attachment, it would be interesting to repeat this probe for other languages. For languages that do not have relative pronoun agreement, one might use singular and plural nouns, and relative clauses where these have the subject role, so the verb can be masked to test whether the model can predict the correctly inflected form.

References

Allein, Liesbeth, Artuur Leeuwenberg, and Marie-Francine Moens (2020), Automatically correcting Dutch pronouns 'die' and 'dat', *Computational Linguistics in the Netherlands Journal* **10**, pp. 19–36. https://www.clinjournal.org/clinj/article/view/102.

Audring, Jenny (2013), A pronominal view of gender agreement, *Language Sciences* **35**, pp. 32–46. https://www.sciencedirect.com/science/article/pii/S0388000112001088.

Bouma, Gosse (2017), Agreement mismatches in Dutch relatives, *Belgian Journal of Linguistics* **31** (1), pp. 136–163, John Benjamins.

Brysbaert, Marc and Don C Mitchell (1996), Modifier attachment in sentence parsing: Evidence from Dutch, *The Quarterly Journal of Experimental Psychology Section A* **49** (3), pp. 664–695, SAGE Publications Sage UK: London, England.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. (2020), Extracting training data from large language models, *arXiv preprint arXiv:2012.07805*.

Conneau, Alexis, 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, *Proceedings of the ACL*, pp. 8440–8451.

Cuetos, Fernando and Don C Mitchell (1988), Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish, *Cognition* **30** (1), pp. 73–105, Elsevier.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), Bertje: A Dutch bert model, *arXiv preprint arXiv:1912.09582*.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: a Dutch RoBERTa-based Language Model, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, pp. 3255–3265. https://www.aclweb.org/anthology/2020.findings-emnlp.292.

Desmet, Timothy, Constantijn De Baecke, Denis Drieghe, Marc Brysbaert, and Wietske Vonk (2006), Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account, *Language and Cognitive Processes* **21** (4), pp. 453–485, Taylor & Francis.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://www.aclweb.org/anthology/N19-1423.

Ettinger, Allyson (2020), What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models, *Transactions of the Association for Computational Linguistics* **8**, pp. 34–48. https://doi.org/10.1162/tacl_a_00298.

Haley, Coleman (2020), This is a bert. now there are several of them. can they generalize to novel words?, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 333–341.

Hemforth, Barbara, Susana Fernandez, Charles Clifton, Lyn Frazier, Lars Konieczny, and Michael Walter (2015), Relative clause attachment in german, english, spanish and french: Effects of position and length, *Lingua* **166**, pp. 43–64. https://www.sciencedirect.com/science/article/pii/S0024384115001722.

Linzen, Tal and Marco Baroni (2021), Syntactic structure from deep learning, *Annual Review of Linguistics* **7**, pp. 195–212, Annual Reviews.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016), Assessing the ability of lstms to learn syntax-sensitive dependencies, *Transactions of the Association for Computational Linguistics* **4**, pp. 521–535, MIT Press.

Manning, Christopher D., Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy (2020), Emergent linguistic structure in artificial neural networks trained by self-supervision, *Proceedings of the National Academy of Sciences* **117** (48), pp. 30046–30054, National Academy of Sciences. https://www.pnas.org/content/117/48/30046.

McCoy, R Thomas, Robert Frank, and Tal Linzen (2020), Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks, *Transactions of the Association for Computational Linguistics* **8**, pp. 125–140, MIT Press.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), Distributed representations of words and phrases and their compositionality, *NIPS*, pp. 3111–3119.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019), How multilingual is multilingual BERT?, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 4996–5001. https://www.aclweb.org/anthology/P19-1493.

Sahin, Gözde Gül, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych (2019), LINSPECTOR: multilingual probing tasks for word representations, *CoRR*. http://arxiv.org/abs/1903.09442.

van Noord, Gertjan (2006), At last parsing is now operational, *in* Mertens, Piet, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42.

Warstadt, Alex, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. (2019), Investigating bert's knowledge of language: Five analysis methods with npis, *arXiv preprint arXiv:1909.02597*.

Wilcox, Ethan, Richard Futrell, and Reoger Levy (2021), Using computational models to test syntactic learnability. https://ling.auf.net/lingbuzz/006327.

Wilcox, Ethan, Roger Levy, Takashi Morita, and Richard Futrell (2018), What do rnn language models learn about filler-gap dependencies?, *arXiv preprint arXiv:1809.00042*.