Computational Linguistics in the Netherlands Journal 14 (2025) 555-582 Submitted 12/2024; Published 07/2025

Evaluating Dutch Speakers and Large Language Models on Standard Dutch: a grammatical Challenge Set based on the *Algemene Nederlandse Spraakkunst*

Julia Pestel* JULIAPESTEL99@GMAIL.COM
Jelke Bloem** J.BLOEM@UVA.NL
Raquel G. Alhama** RGALHAMA@UVA.NL

* *Cognition, Language and Communication BSc Program, University of Amsterdam*

** *Institute for Logic, Language and Computation; University of Amsterdam*

Abstract

This study evaluates the linguistic knowledge of Dutch Large Language Models (LLMs) by introducing a novel challenge set based on the *Algemene Nederlandse Spraakkunst* (ANS). The ANS is a comprehensive resource of Dutch prescriptive grammar created by linguists. We collect acceptability judgements of Dutch native speakers on our dataset, validating its usability while observing varying degrees of grammatical acceptability on specific syntactic phenomena. We evaluate both transformer-encoder and transformer-decoder Dutch LLMs on this dataset, and we compare their performance against the standard rules of Dutch in our dataset and the speaker ratings. We find that transformer-encoder models exhibit almost perfect accuracy on our dataset, yet sensitivities for specific sentences differ between models and humans, partially due to mismatches between the reference grammar and actual use of Dutch.

1. Introduction

The rapidly developing field of machine learning has seen a staggering increase in the use of Large Language Models (LLMs) such as ChatGPT (OpenAI et al. 2024) across a range of disciplines, from applied science to humanities (Bacon 2020). LLMs are utilised for various purposes, including summarising literature, brainstorming new ideas, and writing papers (Bin-Nashwan et al. 2023). These models generate text responses by recognizing linguistic patterns and predicting words from context, based purely on statistical computations (Schwartz et al. 2024). Recent work has been concerned with the evaluation of LLMs on their linguistic knowledge in an effort to determine not only how well they perform on different Natural Language Processing (NLP) tasks, but also whether their performance can be attributed to its internalization of linguistically relevant structures.

Warstadt et al. (2020) introduced a challenge set for evaluating language models on their linguistic knowledge of grammatical phenomena in English. This challenge set, known as the Benchmark of Linguistic Minimal Pairs (BLiMP), consists of *minimal pairs* automatically generated from grammar templates designed by linguists. In this context, minimal pairs are pairs of sentences which are almost identical, but there is a small difference that is sufficient to create a contrast in grammatical acceptability (see Figure 1). Each data set consists of minimal pairs that represent a unique paradigm that isolates contrasts in English morphology, syntax, and semantics. These paradigms are further grouped into one of 12 phenomena; for instance, *Argument Structure* is the phenomenon including paradigms such as *transitive* and *causative* sentences. However, BLiMP focuses exclusively on English, leaving a significant gap in the evaluation of LLMs for other languages. In an effort to bridge this gap, research addressing the grammatical abilities of LLMs in other languages is flourishing. This is also the case for the Dutch language (de Vries et al. 2019, Vanroy 2024), although only BLiMP-NL (Suijkerbuijk et al. 2024) uses minimal pairs.

		Original Dutch minimal pair	English translation
Grammatical	→	1. Het huis is behoorlijk groot.	1. The house is fairly large.
Ungrammatical	→	2. Behoorlijk is het huis groot	2. Fairly is the house large.
		(a)	(b)

Figure 1: An example of minimal pairs where (a) is an example of a Dutch minimal pair used in the challenge set developed in this study and (b) is its English translation. The sentences in the pair differ minimally in word order, but this change is sufficient to create a contrast in grammatical acceptability in Dutch.

The present study aims to contribute to these efforts by developing a set of minimal pairs that complements BLiMP-NL, but differs in a number of ways. Firstly, BLiMP-NL's minimal pairs are based on the *Syntax of Dutch* series of books (Broekhuis 2012, Corver et al. 2015). This is a rather comprehensive and detailed resource aimed at linguists; however, it is written from the perspective of generative syntactic theory. Therefore, many of the minimal pairs involve details relevant to that theory and the categories present in BLiMP-NL are sometimes theory-specific, such as 15. *Parasitic Gaps* and 11. *wh-movement*. Conversely, our minimal pairs are based on the *Algemene Nederlandse Spraakkunst* (ANS, Colleman et al. (2021)), a reference work aimed at the general public. While no linguistic work can be claimed to be theory-neutral, the ANS avoids such theory-specific terminology and categorizations, claiming to "reflect the consensus in the field as much as possible" (Colleman et al. 2021, Section 2.4). This may make our findings of broader relevance to e.g. functionalist linguists.

Secondly, BLiMP-NL makes extensive use of ChatGPT-generated test items (10% human, 90% AI). While we also use such examples to supplement our dataset, we use as many examples from the ANS per category as there are available, resulting in 61% of items being human. Another difference is that Suijkerbuijk et al. (2024) have three exclusion criteria for minimal pairs, ensuring that differences are really minimal.[1] Our study does not use these criteria, therefore we also include structural pairs that differ in terms of larger structural elements, such as types of subordinate clauses.

We evaluate a total of 12 Dutch LLMs on their ability to assign a higher probability to the grammatically acceptable sentence in the minimal pair, as determined in our Challenge Set. In this way, we can determine where the LLMs' sensitivities lie in regard to grammatical distinctions related to each syntactic phenomenon. We compare the performance of the models to grammaticality judgements elicited from Dutch native speakers and weigh to what extent they reflect the same variance in grammatical acceptance of specific sentences.

Thanks to marrying a) a challenge set constructed from a Dutch reference grammar, b) grammatical acceptability judgements of Dutch native speakers and c) LLM simulations, our approach provides insights into the extent to which LLMs have learnt standard grammatical rules of Dutch and to what extent those correlate to human grammatical acceptability judgements.

2. Background

2.1 Large Language Models

In recent years, we have been witness to the rapid development of large language models (LLMs) and have seen major improvement in the performance of these models across a range of Natural Language Processing (NLP) tasks. These developments are partly due to their transformer-based architecture

1. BLiMP-NL uses the following criteria (direct quotation): "(1) the critical region must be the same for the sentences of the minimal pair; (2) at least 1 word directly preceding the critical region must be identical in the sentences of the minimal pair, and (3) the number of words before the critical region can differ by a maximum of 1 word between the sentences of the minimal pair".

(Vaswani et al. 2017). Transformer models are highly flexible neural networks which incorporate an *attention* mechanism which weighs the importance of different words in a sentence. They have the ability to capture complex dependencies and generate high-quality text (Nyandwi 2023). There is a distinction to be made between transformer-decoder and transformer-encoder LLMs.

Transformer-decoder models are typically trained using causal language modelling (CLM) and are therefore also referred to as causal language models. The objective of CLM is predicting the next word in a sequence based on the preceding words, processing text in a unidirectional manner. A notable example is the GPT model family (Kalyan 2024). Transformer-encoder models, on the other hand, are commonly trained using masked language modelling (MLM), and are therefore also referred to as masked language models. The MLM objective allows the model to learn from bidirectional context by masking certain tokens in the input sentence and training the model to predict these masked tokens based on the surrounding context. The BERT family of models popularized the use of bidirectional transformer-encoders with the MLM objective (Devlin et al. 2019). Although CLM seems to be replacing MLM given its potential for language generation, MLM is still a strong contender for tasks that require context awareness, particularly classification (Micheletti et al. 2024, Min et al. 2023).

There are often two stages in the development of LLMs. The first phase is pre-training: this involves training the model (using either CLM or MLM objective) on a large corpus of text data to learn general language representations. The next phase is fine-tuning: this phase trains the model for specific tasks, such as sentiment analysis or entity recognition, normally by leveraging a labelled dataset that is directly relevant to the task. The combination of pre-training and fine-tuning is a powerful approach in NLP, with pre-training providing a strong foundation by learning general language representations and fine-tuning refining the model to specific tasks.

2.2 The Dutch Language in Large Language Models

Currently, large language models are mostly trained on English-language data partially due to the international status that English has and due to the fact that many other languages have fewer training datasets available. These models are often multilingual models: they can understand and provide responses in a variety of languages. However, in most cases the performance of these models decreases when used in a language other than English (Zhu et al. 2023, Papadimitriou et al. 2023, Vlantis and Bloem 2025). This is due to the availability of more training data, more types of annotated data and more interest in conducting research and building systems for English. There remains a significant gap in the development of specialised language models for languages that are not as widely represented as English, such as Dutch. While Dutch has been called a mid-resource language (Kruit 2023) or even a high-resource language, and shares many typological similarities with English, it certainly lags behind English in terms of available models and resources. For example, a Dutch word similarity benchmark was only released recently (Brans and Bloem 2024), and this benchmark showed that Dutch contextual embedding models lag behind their English-language counterparts in terms of accurately representing semantic similarity. In recent years, there have been initiatives to pre-train and/or fine tune LLMs for Dutch NLP tasks (Vanroy 2023, de Vries et al. 2023). The models for these newly developed Dutch LLMs are based on the aforementioned Transformer architectures.

BERTje (de Vries et al. 2019) is an instance of a BERT model (Vaswani et al. 2017) pre-trained on large corpora of solely Dutch texts, making it adept at various Dutch NLP benchmarks including named entity recognition, part-of-speech tagging, and sentiment analysis. The RobBERT model (Delobelle et al. 2020) has also been adapted to Dutch: RobBERT was pre-trained on a much larger web-based corpus than BERTje (39GB of Dutch text against the 12Gb used in BERTje). The large version of RobBERT is trained on the same data for Dutch, but its larger number of parameters endows the model with a more expressive capacity that should be reflected in capturing

longer-distant dependencies. XLM-RoBERTa, a larger multilingual, is trained on the CommonCrawl dataset (Liu et al. 2019), which includes Dutch among its languages.

Among the transformer-decoder models, multilingual OpenAI's GPT-2 (Radford et al. 2019) and Meta's Llama-2 (Touvron et al. 2023) have been used for the development of monolingual Dutch models. The models include GPT-2 large (Havinga 2024) and Llama 2 13B (Vanroy 2023). Both GPT-2 large and Llama 2 13B were pre-trained on a cleaned version of the Dutch part (277GB) of a multilingual dataset (Raffel et al. 2020). The transformer-decoder architecture used in the English Mistral 7B (Jiang et al. 2023) has been used as architecture for the Dutch GEITje models (Rijgersberg 2023, Vanroy 2024). The GEITje models adapt Mistral 7B to Dutch through extensive pre-training on Dutch corpora, such as Dutch Gigacorpus[2] and MADLAD-400 (Kudugunta et al. 2023, Vanroy 2023), supplemented by task-specific datasets to refine performance in specific applications.

2.3 LLM Evaluation: Benchmarks of Minimal Pairs

There are a multitude of motivations for studying the scope of LLMs' linguistic knowledge, one of which is the evaluation of these models on how well they capture linguistic phenomena (Bacon 2020). Determining which sensitivities LLMs have to different phenomena increases the confidence in the abilities and output of these models. This information can in turn be useful for the optimisation of the models. Furthermore, evaluations help compare models for their suitability for varying functionalities and use cases (Bacon 2020, Belinkov and Glass 2019).

In earlier years, evaluation of the linguistic capabilities of language models was focused on the quality of their semantic representations, using word similarity or sentence similarity benchmarks containing human-rated pairs of items, the scores of which should correlate to model scores for the same items. These correlations are considered to reflect model quality and to predict a model's performance on downstream tasks to some extent. This type of intrinsic evaluation contrasts with the extrinsic approach to evaluating models, where performance on a downstream task is evaluated. Intrinsic evaluation of language models is often done with reference to human-rated benchmarks or with post-hoc evaluations of model outputs by human raters. A range of methods has been proposed to evaluate the quality of language models intrinsically. Bakarov (2018) identifies sixteen of them. A prominent example of this is the SimLex-999 semantic similarity dataset (Hill et al. 2015), created by instructing participants explicitly about the concept of semantic similarity. This focus on consciously or expertly elicited judgements has remained popular, but subsequently, interest shifted to grammaticality rather than semantics. There was also increased interest in probing for representations of specific linguistic phenomena, such as negative polarity (Bylinina and Tikhonov 2022) or Construction Grammar constructions (Veenboer and Bloem 2023).

One of the most influential grammaticality benchmarks was pioneered by Warstadt et al. (2020) for evaluating the linguistic knowledge of LLMs on major grammatical phenomena in English. The Benchmark of Linguistic Minimal Pairs (shortened to BLiMP) consists of 67 individual datasets with each dataset containing 1000 minimal pairs which were generated using linguist-crafted grammar templates. These minimal pairs were manually verified by human validators through a forced-choice task. The validators were presented with a subset of minimal pairs from BLiMP and were asked to determine which of the two sentences in the pair was grammatical.

With BLiMP, the authors evaluated different types of LLMs, namely GPT-2 and Transformer-XL, and compared them to baselines such as a Long Short Term Memory model (Hochreiter and Schmidhuber 1997) and n-gram models. The models were evaluated on grammatical phenomena concerning morphology (e.g., anaphor agreement), syntax (e.g., argument structure) and semantics (e.g., control/raising). The LLMs were evaluated on BLiMP by calculating the proportion of the 67.000 minimal pairs where a model assigned a higher probability to the grammatical sentence over the ungrammatical sentence. The authors found that all LLMs under-performed when compared to

2. http://gigacorpus.nl/

human accuracy. Of the LLMs, GPT-2 had the highest accuracy, while the 5-gram model had the lowest and, also on the low side, Transformer-XL and the LSTM model had similar accuracies to each other.

After Warstadt et al. (2020) introduced BLiMP, datasets inspired by this benchmark have been developed to evaluate the linguistic knowledge of LLMs in other languages. These benchmarks include a Japanese challenge set (JBLiMP, Someya and Oseki (2023)), and challenge sets CLiMP (Xiang et al. 2021) and SLING (Song et al. 2022) for Mandarin Chinese. Most relevant to our work, recently Suijkerbuijk et al. (2024) developed a benchmark for the evaluation of the linguistic knowledge of LLMs in Dutch, named BLiMP-NL.

GPT-2 (Radford et al. 2019) turned out to be the best-performing model in JBLiMP and BLiMP-NL, although in the latter it was on par with BERTje (de Vries et al. 2019), a BERT-based model. While the former study only compared two versions of GPT-2 models and two non-Transformer architectures (namely LSTM and n-grams), the latter also evaluated a range of Transformer-based models, including both transformer-encoders and transformer-decoders. In CLIMP, Chinese BERT (Devlin et al. 2019), a BERT-based model, emerged as the best-performing model, compared to non-Transformer models (again, LSTMs and n-grams). When comparing models to human acceptability judgements, Chinese BERT appears to be closest to human performance in CLiMP, while all the models in JBLiMP fall short. In the case of BLiMP-NL, only GPT-2 was compared to human performance but the authors did not find evidence of comparable performance.

The generation methods in these datasets, however, pose a source of variation. The original BLiMP was developed using grammar templates and a vocabulary to automatically generate 1000 minimal pairs per paradigm. The same method was used by Xiang et al. (2021) for their dataset. Song et al. (2022) created their own grammar templates, with the goal of generating more natural minimal pairs, since those in the Xiang et al. (2021) dataset were occasionally nonsensical or appeared artificial. Employing a completely different method, Someya and Oseki (2023) did not use grammar templates but instead extracted ungrammatical sentences and their grammatical counterparts from syntax journal articles. If they could not find a grammatical counterpart, they would manually create one. Utilising this alternative method, the authors were able to include more complex phenomena in their meticulously developed dataset consisting of 331 minimal pairs. Suijkerbuijk et al. (2024) also diverged from the aforementioned methods for the generation of minimal pairs by opting not to use grammar templates or extract sentences. Instead, they manually created 10 minimal pairs for each paradigm and utilised ChatGPT to extend the dataset to 100 minimal pairs per paradigm. The pairs generated by ChatGPT were manually checked to confirm they were sensical and targetted the intended grammatical phenomena.

These studies also report human performance on the created datasets, generally with the goal of (1) validating the datasets, ensuring that the grammatical contrasts are known to language users, or (2) comparing the performance of the models to that of humans, when sufficient variability is present in the human responses. In the case of BLiMP-NL, Suijkerbuijk et al. (2024) opted to retrieve acceptability judgements from their validators on a Likert scale instead, ranging from 1 to 7. This decision was made seeing as they found it important to test whether they can show sensitivity to this acceptability as acceptability judgements are not binary in nature, but gradient.

3. Creating the Challenge Set

Our dataset is based on the *Algemene Nederlandse Spraakkunst* (ANS): the reference work for all aspects of Dutch grammar (Colleman et al. 2021). We used the e-ANS, the ANS website[3], to retrieve the minimal pairs, by employing web scraping. In the following, we introduce the ANS resource, the

3. `https://e-ans.ivdnt.org/`

syntactic phenomena we focus on, and the procedure to extend the sets of retrieved minimal pairs with additional pairs.[4]

3.1 The ANS

The ANS (*Algemene Nederlandse Spraakkunst*) is the authoritative reference grammar for the Dutch language, created by a committee of Dutch and Belgian linguists who first published it as a book in 1984 (Haeseryn et al. 1984). The ANS aims to provide as complete a description as possible of the grammatical aspects and rules of contemporary Standard Dutch. Standard Dutch is defined as the language that is generally used in public communication, i.e. in all major sectors of public life: education, jurisdiction, media, administration, etcetera.[5] The ANS caters to both experienced users of Dutch, such as students and teachers at the academic level, and individuals with sufficient basic knowledge to be able to use an extensive Dutch grammar.

The ANS consists of four large sections: 'The Sound', 'The Word', 'The Constituent', 'The Sentence', and 'General Phenomena'. The first section, 'The Sound', covers Dutch phonetics and phonology. The section covering 'The Word' contains a general introduction and a description of the traditionally distinguished word types. The information in this section is limited to morphological matters. The remaining sections pertain to syntactic aspects of Dutch grammar. In the third section, 'The Constituent', a description is given of the possibilities of combining words into constituents such as noun phrases and verb phrases, and of the mutual order of these words within such phrases. The fourth section pertains to 'The Sentence', where the syntax of the sentence is addressed in addition to the order of sentence constituents, traditional sentence parts, and various types of sentences. The last section, 'General Phenomena', deals with phenomena that play a role at different levels of grammar, such as juxtaposition and negation, contraction, modality, etcetera.

One of the resources of the ANS is a compilation of representative examples of Dutch language. While some of these are in the form of a single sentence which substantiates a phenomenon of Dutch grammar, other examples consist of two minimally different sentences where one of the sentences has been assigned a label.[6] The most common labels used in these types of examples are as follows: *informeel* (informal), *formeel* (formal), *uitgesloten* (ruled-out) and *twijfelachtig* (questionable). Furthermore, the ANS uses labels to indicate regional variation; examples are *in NN* (in Netherlandic Dutch), *in BN* (in Belgian Dutch), and *in SN* (in Surinamese Dutch).[7] We are interested in the examples where one of the sentences is labelled as being *uitgesloten* (ruled-out) and has no label indicating any kind of regional variation. These examples of minimally different sentences that vary in grammatical acceptability are the minimal pairs that we retrieve for our dataset.

By employing web scraping, we extracted the minimal pairs from our chosen section of the e-ANS along with their labels. Below is an example minimal pair from the nominalizations category of the dataset, of which the first is ungrammatical according to ANS:

(1) * Het grootmoeder inschenken van thee (is een hele opgave.)
The grandmother pouring of tea (is a whole task.)

(2) Het voor grootmoeder thee inschenken (is een hele opgave.)
The for grandmother tea pouring (is a whole task.)
'Pouring tea for grandmother is quite a task'.[8]

This process resulted in a dataset consisting of 631 minimal pairs across the linguistic phenomena, which we later extended (as explained below).

4. Our dataset and raw results can be found at: `https://github.com/juliapestel/evaluating-dutch-LLMs`
5. Source: `https://e-ans.ivdnt.org/over`: 5.1 Het concept 'standaardtaal'
6. Note that the ANS does not follow strict criteria for what constitutes a 'minimal' difference.
7. An overview of the labelling system used by the ANS can be found at `https://e-ans.ivdnt.org/over#ans000403st`: 5.5 Het gehanteerde labelsysteem
8. Example from `https://e-ans.ivdnt.org/topics/pid/ans140803lingtopic`

3.2 Coverage of Syntactic Phenomena

Our goal was to create a challenge set that was as complete and reliable as possible for one of the five sections, namely: the CONSTITUENT. This decision was made in part due to this section having the largest amount of minimal pairs after extraction (207 minimal pairs). Additionally, we found that most related work evaluates the linguistic knowledge of language models by looking at general linguistic phenomena, as is done in BLiMP and BLiMP-NL, and not on constituent types specifically.

The section the CONSTITUENT is further divided into five different types of constituents; *De naamwoordelijke constituent* (The Nominal CONSTITUENT), *De adjectivische constituent* (The Adjectival CONSTITUENT), *De bijwoordelijke constituent* (The Adverbial CONSTITUENT), *De adpositieconstituent* (The Adpositional CONSTITUENT) and *De werkwoordelijke (verbale) constituent* (The Verbal CONSTITUENT). Each of these CONSTITUENT types cover varying syntactic phenomena. However, the e-ANS does not provide minimal pair sentences for some of these phenomena. We chose to limit the challenge set to cover the syntactic phenomena that have at least one minimal pair as defined on the e-ANS. Our resulting dataset includes the following 10 phenomena:

1. **Definite, Indefinite, Categorical and Generic Nominal Constituents** (e-ANS 14.3: *Bepaalde, onbepaalde, categoriale en generieke naamwoordelijke constituenten*): The distinction between definite, indefinite, categorical and generic nominal constituents. A definite nominal constituent is a CONSTITUENT that designates one or more identified selves. With indefinite constituents we introduce persons or things. In categorical and generic nominal constituents the head abstracts from individual cases with categorical referring to a category of class and generic referring to a generalisation of a species or class.

2. **Nominalizations** (e-ANS 14.8: *Nominalisaties*): Where the nominal constituent can be related to a corresponding (active or passive) sentence.

3. **The Construction of the Adjective Constituent** (e-ANS 15.2: *De bouw van de adjectivische constituent*): This covers the construction of the adjectival constituent where the CONSTITUENT has an adjective as its head. This can be preceded by modifying adverbs or followed by modifying prepositional constituents.

4. **Degree-Indicating or Reinforcing (Pre)Determinations** (e-ANS 15.3.1.1 *Graadaanduidende of versterkende (voor)bepalingen*): To indicate the intensity of the property or state expressed by an adjective (degree designation) or to indicate that the said property or state is highly valid (reinforcement), constituents can be used especially with an adverb or adjective at its core in adverbial function.

5. **Subordinate Clauses** (e-ANS 15.4.3: *Bijzinnen*): This covers the subordinate clause; a clause that cannot stand alone as a complete sentence as it merely complements a sentence's main clause, thereby adding to the whole unit of meaning.

6. **Pronominal Adverbs** (e-ANS 17.1.2: *Voornaamwoordelijke bijwoorden*): This covers the use of the pronominal adverb; a combination of a preposition and a pronoun (i.e. *hier* + *bij* = *hierbij*, hereby)

7. **Group-Forming and Non-Group-Forming Uses of Verbs** (e-ANS 18.5.1.2: *Groepsvormend en niet-groepsvormend gebruik van werkwoorden*): This covers the use of group-forming and non-group-forming verbs where group-forming verbs are verbs that connect to another verb and change or add something to the meaning of the other verb while also determining its form.

8. **Verbs With a Participle as a Complement** (e-ANS 18.5.2: *Werkwoorden met een deelwoord als aanvulling*): This covers the addition of a participle as a complement to a verb in a

verbal constituent. The participle is a nonfinite verb form that has some of the characteristics and functions of both verbs and adjectives.

9. **Verbs With an Infinitive as a Complement** (e-ANS 18.5.4: *Werkwoorden met een infinitief als aanvulling*): This covers the use of a complementary infinitive: an infinitive used with a verb whose meaning is not felt to be complete.

10. **The Order Within the Verb Ending Group** (e-ANS 18.5.7: *De volgorde binnen de werkwoordelijke eindgroep*): This covers the order of the verbs within a verbal CONSTITUENT focussing on bipartite end groups and the verb infinitive (also known as verb clusters).

The chapters of the ANS2 version (1997/2002) are currently undergoing extensive revision, which is expected to last until 2027/2028.[9] Seeing as this revision project is ongoing, some of the chapters which cover the syntactic phenomena we cover in the challenge set have not yet been revised. Only the chapter covering The Adpositional CONSTITUENT has recently been revised, the other chapters (The Nominal CONSTITUENT, The Adjectival CONSTITUENT, The Adverbial CONSTITUENT and The Verbal CONSTITUENT) are still subject to revision.

The main difference between ANS2 (1997/2002) and ANS3 (newly revised) of The Adverbial CONSTITUENT is that the former is far less extensive; the grammar rules in the new version are explained in more detail and substantiated with examples. Additionally, the new version contains more sub-chapters, creating a more in-depth overview of The Adverbial CONSTITUENT and its grammar rules. When ANS3 is done, it would be worth revising our dataset with any new minimal pairs added.

3.3 Generating Additional Minimal Pairs

To construct the dataset, we retrieved the minimal pairs provided on the ANS website; the e-ANS. The available amount of pairs ranged from 1 to 49 pairs per phenomenon, and we extended each set with additional pairs, until reaching 50.

The additional minimal pairs were generated using the pairs extracted from the e-ANS as a guide, so that the generated minimal pairs varied with regard to the same grammatical aspect. The additional minimal pairs were crafted manually or generated using the ChatGPT 4.0 generative language model (OpenAI et al. 2024). Table 1 shows an overview of the sources of minimal pairs for each phenomenon. Examples of the minimal pairs for each syntactic phenomena are provided in Table 2.

Phenomena	ANS	Manual/ANS	Manual	Chat-GPT	Total
1. Definite, Indefinite, Categorical and Generic Nominal Constituents	11	0	18	21	50
2. Nominalizations	15	0	13	22	50
3. The Construction of the Adjective Constituent	1	3	27	19	50
4. Degree-Indicating or Reinforcing (Pre)Determinations	1	3	27	19	50
5. Subordinate Clauses	2	0	9	39	50
6. Pronominal Adverbs	29	5	16	0	50
7. Group-Forming and Non-Group-Forming Uses of Verbs	8	0	16	26	50
8. Verbs With a Participle as a Complement	11	0	4	35	50
9. Verbs With an Infinitive as a Complement	49	1	0	0	50
10. The Order Within the Verb Ending Group	15	2	19	14	50
Grand Total	**142**	**14**	**149**	**195**	**500**

Table 1: *Generation method per phenomenon. The column 'Manual/ANS' refers to minimal pairs that were slightly modified due to their incomplete extraction from the e-ANS.*

All the minimal pairs are based on the pairs extracted from the e-ANS. In the rare cases where one or both of the extracted sentences in a minimal pair were incomplete, we modified them slightly

9. Timeline of ANS chapter revisions: `https://e-ans.ivdnt.org/qanda#herzienehfds`

to ensure comprehension. Often, these irregularities in the minimal pairs were due to the irregular way the pairs were presented on the e-ANS, which caused our Web Scraper to extract incorrect information. Typically, the ANS presents a minimal pair as a numbered example consisting of two sentences; *a*, the grammatical sentence, and *b*, the ungrammatical sentence. Atypically, the ANS presents both sentence *a* and *b* in an example as grammatically correct sentences, and presents the ungrammatical counterparts of these sentences in a separate example.[10] We reviewed these by consulting the webpage from which the sentences were extracted and minimally modified the minimal pairs to ensure an accurate representation of the syntactic phenomenon. Of the 500 minimal pairs across all phenomena and the 156 pairs extracted from the e-ANS, 14 pairs were minimally altered. Seeing as these minimal pairs mostly consist of information from the e-ANS but needed minor manual revision to complete them, we class their generation under 'Manual/ANS'. An overview of these 14 minimal pairs can be found in Appendix B.

For the generation of minimal pairs using ChatGPT, we provided the model with the minimal pairs extracted from the e-ANS for a specific syntactic phenomenon and prompted it to create additional pairs that had the same grammatical variation between them as in the examples. In the instances where we felt ChatGPT needed more examples than provided by the ANS to generate additional pairs that accurately represent the syntactic phenomena, we supplied the model with the 'Manual/ANS' generated pairs and/or further hand-crafted minimal pairs. All generated minimal pairs, whether handcrafted or generated using Chat-GPT, were manually checked by a native Dutch speaker to ensure that they correctly represent the intended syntactic phenomenon and did not contain any additional errors.

Constituent	*Ph.*	*Grammatical sentence*	*Ungrammatical sentence*
Noun	1	De maan schijnt. *The moon shines.*	Er schijnt **een maan**. ***A moon** is shining.*
	2	Het bakken van pannenkoeken (is voor Anneke niet moeilijk.) *Baking pancakes (is not difficult for Anneke.)*	**Het van** pannenkoeken bakken (is voor Anneke niet moeilijk.) ***Pancakes Baking** (is not difficult for Anneke.)**
Adjective	3	Een zeer hard gesteente. *A very hard rock.*	Een **zeer gesteente**. *A **very rock**.*
	4	De koffie is warm genoeg. *The coffee is warm enough.*	Warm **genoege** koffie. *Hot **enough** coffee.**
	5	(Jan is altijd) bereid (om) te helpen. *(Jan is always) willing (to) help*	(Jan is altijd) bereid **dat hij zou** helpen. *(Jan is always) willing **that he would** help.*
Adposition	6	Je moet tenslotte ergens in geloven. *You have to believe in something after all.*	**Ergens in moet** je tenslotte geloven. ***Something in have you** to believe, after all.**
Verb	7	Hij zei dat hij de kraanvogels graag wilde fotograferen. *He said he would like to photograph the cranes.*	Hij zei dat hij de kraanvogels **wilde graag** fotograferen. *He said he **like to would** photograph the cranes.**
	8	Zijn broer gaat in Spanje wonen. *His brother is going to live in Spain.*	Zijn broer is in Spanje **gegaan wonen**. *His brother is to live in Spain **went**.**
	9	Wim zit te slapen. *Wim is sleeping.*	Wim **zit slapen**. *Wim **sleeping**.**
	10	Hij zei dat hij het vliegtuig niet kon zien naderen. *He said he couldn't see the plane approaching.*	Hij zei dat hij het vliegtuig **niet naderen kon** zien. *He said he **couldn't the plane approaching see**.**

Table 2: *Example minimal pairs for all 10 phenomena (Ph., see description in Table 1), with English glosses. For clarity, the words that determine the ungrammaticality of the sentence are in bold for both the example and its English translation. As these are translations rather than glosses, in some cases the English translation of the minimal pair is not grammatically identical to the original: these are indicated with an asterisk (*).*

4. Acceptability Judgements of Dutch Native Speakers

We evaluated adult native speakers of Dutch on a subset of our challenge set. Grammatical reference works are often written with a binary notion of grammaticality in mind. This is a simplifying

10. For an example in the ANS of this atypical notation of a minimal pair, please refer to: `https://e-ans.ivdnt.org/topics/pid/ans140302lingtopic` (14.3.2.2)

assumption that was commonly made in works on syntactic theory, though it is the subject of a long-standing debate and it is increasingly being questioned.

Already raised by Chomsky (1965, p. 152-153), who argued for a binary notion of grammaticality, the question often involved discussion regarding the boundaries of syntax and other mechanisms, such as semantics, pragmatics or general cognitive mechanisms affected by input frequency. In binary accounts, uncertainty in acceptability judgements was attributed to these other factors (e.g. semantic factors by Chomsky (1965)), while gradient accounts claim that there are gradient aspects to grammatical competence that cannot be explained by such other factors (Lau et al. 2017). Generative syntactic theory often adheres to a binary view, but proposals have also been raised to specify the role of non-syntactic factors such as frequency in grammar as opposed to production. An example of this is the proposal of Featherston (2005), who theorizes an output selection module that is affected by the factor of frequency. In usage-based linguistics, frequency is more central to the development of grammatical structure (Divjak 2017), leading to theoretical proposals of probabilistic mental representations of grammatical patterns (Bod et al. 2003). An extensive overview of how different lines of linguistic theory view this debate is presented by Francis (2022).

Even if one assumes that grammaticality is binary for an individual language user, acceptability judgements only provide an approximation of it, and grammaticality measured in aggregate over a population may exhibit variation. Therefore, to provide an alternative to the labels from ANS and to determine the extent of the agreement between the grammatical contrasts in our dataset and actual acceptability judgements of Dutch users, we present a behavioural experiment in which we use our dataset to elicit responses on grammatical acceptance for our sentences, from Dutch native speakers.

4.1 Design

In order to effectively and efficiently test native speakers, we created a subset of the challenge set with nine minimal pairs per phenomenon. The nine minimal pairs for each of the phenomena were chosen based on how they were created: we first selected the minimal pairs which were originally from the ANS and, when insufficient, the subset was extended to nine pairs, prioritizing the hand-crafted pairs over those generated by ChatGPT.

Following Suijkerbuijk et al. (2024), we divided the ten phenomena evenly over five experiments, with each experiment testing two different phenomena. Each experiment was further divided into three versions. Each of these versions contained six grammatical sentences and three ungrammatical sentences for both phenomena tested in that experiment. One version never contained both sentences from a minimal pair: these were always divided between versions. With three versions for each of the five experiments we created a total of 15 different surveys with each survey containing 18 sentences: nine from both phenomena. While Suijkerbuijk et al. (2024) used 108 sentences per participant, we opted for 18 sentences per participant in our study to prevent fatigue effects near the end of the experiment.

The order in which the sentences were displayed to the participant was randomised. The experiments were carried out using web-based surveys which were created using Qualtrics software (Qualtrics 2005). Using the web application Nimble Links[11], we created a link that randomly assigned participants to one of the 15 surveys.

4.2 Participants

Participants qualified to take part in the experiment if they were a native speaker of Dutch, currently lived in the Netherlands and did not have any language disorders such as dyslexia. We recruited 132 participants by sharing the survey link with different individuals who met the criteria. This link was shared using WhatsApp and the social media platforms Instagram, LinkedIn, and Facebook.

11. `https://www.nimblelinks.com/`

The recruitment message included a brief (non-revealing) description of the study and a link to the survey. Recruitment was limited to participants who resided in the Netherlands for practical reasons.

4.3 Procedure

Participants were assigned a survey by clicking on the link that was shared with them. Hereafter they were presented with the information brochure and consent form. If the participant consented to the terms, they were directed to the next page where they were asked a series of questions to confirm whether they met the criteria to participate. The participant was asked whether Dutch is their native language, if they currently reside in the Netherlands, and if they do not have a language disorder. When the questions were answered in the affirmative, the participant was directed to the next page where they were shown instructions. The instructions described the process of the experiment: the participant would be presented with a list of sentences and asked to rate them on a Likert scale (Likert 1932), from 1 (very bad) to 7 (very good), by moving a slider placed directly under each individual sentence. Like Suijkerbuijk et al. (2024), we opt for a Likert scale as opposed to a forced-choice task to retrieve graded acceptability judgements. In the instructions it was additionally specified that when deciding on a rating for the sentence, only the grammatical acceptability should be considered, not the spelling or the vocabulary of the sentence as these are correct in all instances. This was mentioned as some words in the sentences obtained from the e-ANS are not commonly used in standard Dutch, which may cause confusion for the participant. Below, an example of such a minimal pair is shown, with the uncommon word being *tuk*. After carefully reading the instructions, the participant could start the experiment by continuing to the next page.

(3) Hij is niet bepaald tuk op vergaderen.
He is not exactly keen on meetings.

(4) * Hij is niet bepaald tuk.
He is not exactly keen.

4.4 Results

The average acceptability judgements per syntactic phenomenon can be found in Figure 2. As we can see, native speakers unsurprisingly assigned a lower score to the ungrammatical sentences than to the grammatical sentences. The overall mean scores for all the grammatical and ungrammatical sentences is 5.52 (1.07) and 2.28 (0.92), respectively. This entails a difference of 1.48 from the maximum score (7) in the case of grammatical sentences, and a difference of 1.28 from the minimum score for ungrammatical sentences. Compared to human participants on BLiMP-NL, we find that participants on our experiments tend to give more differentiated scores, which may suggest that grammaticality distinctions are clearer on our dataset –although other variables arising from differences in conducting the experiment may also be responsible for this. In any case, the data suggests that our dataset is broadly in line with the grammatical intuitions of Dutch native speakers, which confirms the usability of our dataset.

When looking at the differences in acceptability ratings across the different types of constituents, we see that native speakers have more clearly differentiated scores between grammatical and ungrammatical sentences in phenomena 3 (The Construction of the Adjective Constituents), 5 (Subordinate Clauses), and 8 (Verbs with participle). Surprisingly, these phenomena belong to different CONSTITUENT categories, but at least two of these cover phenomena that are frequent, clear and quite localized in a constituent, involving short example sentences. This is presumably easier to evaluate for grammaticality. The ungrammatical sentences for phenomenon 5 (Subordinate Clauses) largely consist of sentences with the wrong kind of relativizer, as in this example:

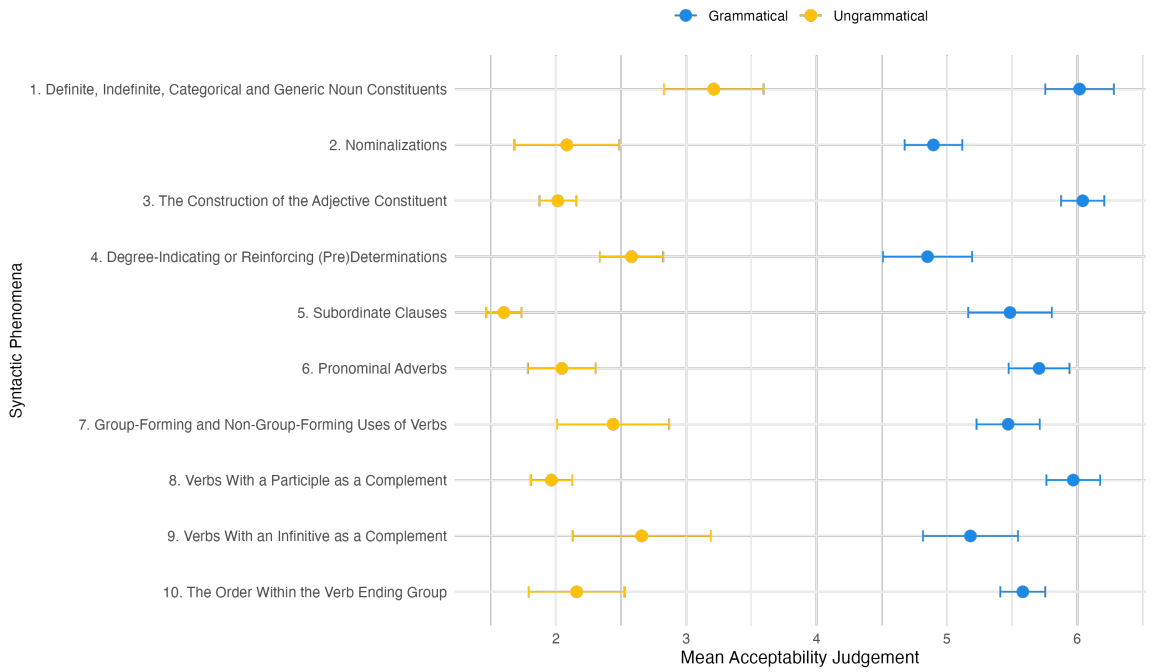


Figure 2: *Mean acceptability ratings for the ungrammatical and grammatical sentences for each phenomenon: Nominal Constituent (1, 2), Adjective Constituent (3, 4, 5), Adposition Constituent (6), Verbal Constituent (7, 8, 9, 10). Grammatical mean: 5.52 (1.07), ungrammatical mean: 2.28 (0.92).*

(5) * (Jan is altijd) bereid dat hij zou helpen.
(John is always) willing that he would help.

(6) (Jan is altijd) bereid (om) te helpen.
(John is always) willing (REL) to help.
'John is always willing to help.'

Subordinate clauses are large constituents and multiple words are changed (this type of minimal pair would have been excluded by Suijkerbuijk et al. (2024)). This leads to ungrammaticality that is salient and easy to detect for participants (6.56 vs 1.15 average grammaticality rating for this pair).

The sentences for phenomenon 3 are diverse, but include many examples where the word that the adjective is supposed to modify is missing in the ungrammatical condition, to demonstrate the required elements of the adjectival constituent:

(7) Wim is nogal eigenwijs.
Wim is rather stubborn.

(8) * Wim is nogal.
Wim is rather.

These are also salient and elicit strong judgements (6.5 vs 2.4 acceptability).
For phenomenon 8, we find examples such as:

(9) We zagen de koningin voorbijkomen.
We saw the queen pass_by.

(10) * We hebben de koningin gezien voorbijkomen.
We have the queen seen pass_by.

This pair shows a difference in aspect that involves adding the auxiliary verb 'hebben', thereby also changing 'zagen' to its past participle 'gezien' and moving it from the verb-second to the verb-final position. This involves two changes, hence this more salient difference may have made it easier for the participants to detect the ungrammaticality (6.7 vs 1.8 average grammaticality).

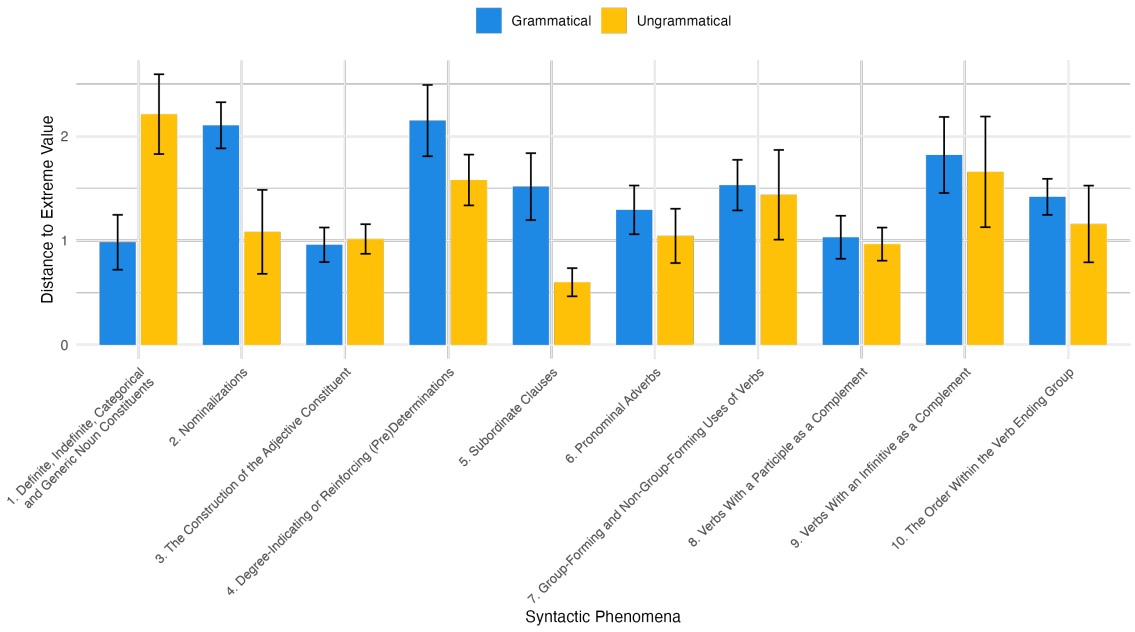


Figure 3: *The difference between the mean acceptability ratings of native speakers and the gold values for each grammatical condition per phenomenon. The gold values for grammatical = 7 and ungrammatical = 1.*

For additional clarity, Figure 3 shows only the difference between the observed scores and the 'gold' score (i.e. 1 for ungrammatical and 7 for grammatical sentences); hence lower bars correspond to sentences whose score is closer to the idealized score based on the e-ANS labels. Here we see that higher bars are for 2 (Nominalizations), 4 (Degree-indicating or Reinforcing (Pre)Determinations), and 9 (Verbs with an infinitive) in the case of grammatical sentences.

Nominalizations include examples such as 1 and 2 mentioned in the introduction. This pair is rated 2.81 and 1.5: even the supposedly grammatical variant receives a low rating. This is probably because nominalizations are infrequently used in Dutch and there are usually more conventional ways to phrase the same sentences that participants would be able to think of. [12]

For phenomenon 4, degree indicators, the grammatical sentences receive somewhat low scores. For example:

(11) Die flats zijn geweldig hoog.
Those flats are incredibly high.

12. Team Taaladvies on nominalizations in Dutch: `https://taaladvies.net/naamwoordstijl-algemeen/`

(12) * Die flats zijn geweldige hoog.
Those flats are incredibly high.

In contrast to many of the phenomena where ratings differed strongly, this minimal pair is very minimal, differing only in the inflection of the adjective. There are various similar examples in this category. Furthermore, the example uses a less common meaning of the adjective 'geweldig', perhaps leading to less familiarity and lower overall acceptability scores among some speakers.

For phenomenon 8, infinitival verbs, the items are somewhat diverse, but quite a few of them involve 'te'-infinitival verbs:

(13) De jongens zitten de hele les te slapen.
The boys sit the whole class to sleep.

(14) * De jongens zitten de hele les slapen.
The boys sit the whole class sleep.
'The boys sleep the entire lesson.'

This pair is rated 6.2 vs 3.6, and while this pair does not seem so controversial to us, the 'te' of te-infinitival verbs is often optional in other contexts, and there are also pairs in the dataset where we think both options would be grammatical to many speakers.

Conversely, some 'grammatical' examples in this category got a rather low rating, such as:

(15) De appels hebben al een maand aan de boom hangen te rotten.

The use of 'hangen' as a grouping verb for an infinitival verb is quite uncommon – only 11 out of 827.709 two-verb clusters studied by Bloem et al. (2017) use this grouping verb. This may explain why our participants rate this sentence with 2.8, while the minimal pair is mainly meant to illustrate a point about the order of the two verbs.

There are quite a few word order variation phenomena around the Dutch verbal constituent. This is especially the case for verb clusters (Coussé et al. 2008), including regional variation (Barbiers et al. 2018), where ongoing changes even cause differences in usage between older and younger speakers (Olthof et al. 2017). It has also been noted that Dutch speakers sometimes have strong prescriptive opinions on verb orders (Swerts and van Wijk 2005), which may affect grammaticality judgements. This extensive optionality probably makes it more difficult for participants to clearly judge such sentences as grammatical or ungrammatical.

It is especially interesting that 9 (verbs with an infinitive as a complement) and 8 (verbs with a participle) show opposite human judgement patterns, as these two categories are very similar. But this seems to be due to the types of examples that are discussed in ANS, with the discussion of infinitives focusing more on possibly optional elements.

Overall the ungrammatical sentences exhibit a smaller difference with the ideal score, with the exception of Pronominal Adverbs (6), for which Dutch speakers deviate over 2 average points from the minimum score. The ungrammatical sentences for this phenomenon largely involve treating unique things as indefinites, as in these examples:

(16) * Een sahara is erg heet.
A sahara is very hot.

(17) * Er schijnt een maan.
There shines a moon.

Both of these are considered ungrammatical by ANS, but perhaps due to the semantic nature of this restriction, some examples were often considered grammatical by participants. It is easy to imagine a science fiction setting in which Example 17 is perfectly felicitous. This supposedly ungrammatical example was rated 5.25 on the 7 point scale on average, and this is the cause of the large difference for this category.

Overall, we observed that native speaker acceptability judgements do not always follow the grammaticality labels assigned by the ANS authors. This can be due to various factors besides grammaticality that affect acceptability judgements such as the possibility to consider prescriptive norms in a survey-style experiment, as extensively discussed in experimental syntax textbooks (Goodall 2021, Sprouse 2023), as well as due to the possibility of grammaticality being a gradient phenomenon or appearing to be gradient due to non-syntactic factors, as discussed at the beginning of this section. In particular, factors such as frequency of words and constructions, the salience of the difference between pairs (e.g. in number of words), and the extent to which constraints are semantic in nature appear to play a role, and can pertain to general cognitive mechanisms or interfaces of syntax.

5. Evaluating LLMs

5.1 Models

Using the challenge set we created, we evaluated the most prominent Dutch LLMs, which are the same that were used in Suijkerbuijk et al. (2024).

Transformer-decoder We evaluate eight transformer-decoder models that are pre-trained using causal language modelling (CLM). Two such models are based on the GPT-2 architecture (Radford et al. 2019); particularly GPT-2 large (Havinga 2024) and GPT-2 small Gro-NLP (de Vries and Nissim 2021). The next two are based on Llama 2-based models (Touvron et al. 2023) which are fine-tuned to perform better in the Dutch language, namely, Llama 13B and Llama 13B chat (Vanroy 2023). The latter four models are based on Mistral (Jiang et al. 2023); particularly GEITje 7B and GEITje-7B-chat (Rijgersberg 2023), GEITje-7B-ultra-sft and GEITje-7B-ultra-dpo (Vanroy 2024).

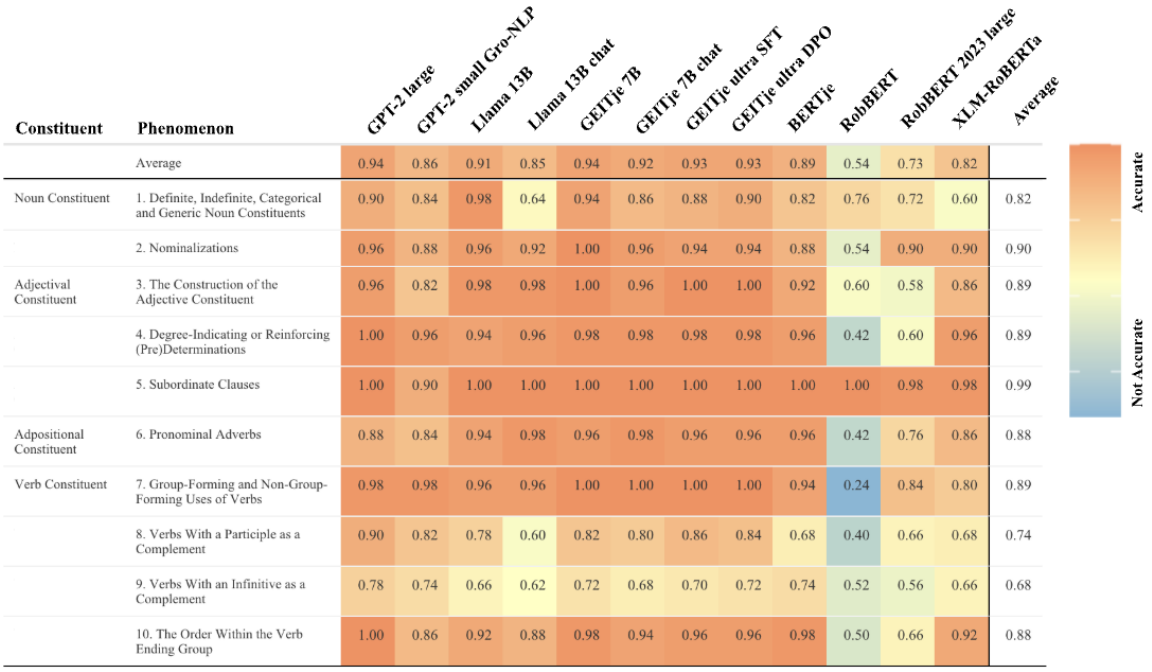
Transformer-encoder This study evaluates four transformer-encoder models that are pre-trained using MLM: BERTje (de Vries et al. 2019), which is based on BERT (Devlin et al. 2019), and three RobBERTa based models (Liu et al. 2019): RobBERT (Delobelle et al. 2020), RobBERT large (Delobelle and Remy 2024), and XLM-RoBERTa (Conneau et al. 2020).

5.2 Evaluation Methods

We evaluate the LLMs based on model output (extrinsically) by employing language model scoring, i.e. obtaining the probabilities the LLM gives to both the grammatically correct and incorrect sentences in a minimal pair, allowing us to directly compare them. The process of scoring causal LLMs (transformer-decoder) differs from scoring masked LLMs (transformer-encoder) due to the differences in modelling. CLM processes text in a unidirectional manner, meaning we obtain the probabilities for causal LLMs by applying the chain rule and summing the log-likelihood values for each successive token. Masked LLMs, on the other hand, learn from bidirectional context, which makes getting the probabilities a more complex process. To effectively and efficiently obtain the probabilities for these models, we follow the method used by Suijkerbuijk et al. (2024) in BLiMP-NL by estimating the pseudo-log-likelihood score using the PLL-word-l2r metric from Kauf and Ivanova (2023). The probabilities are normalised by sentence length for both causal and masked LLMs.

5.3 Performance of Large Language Models

After obtaining the probabilities of both the grammatically correct and incorrect sentences in each minimal pair, we calculated the accuracy of the model by determining the proportion of minimal pairs for which the LLM assigned a higher probability to the grammatically correct sentence. In addition to the overall accuracy of the LLMs, we calculated the accuracy of the LLMs for each syntactic phenomena. The results can be found in Figure 4.



Constituent	Phenomenon	GPT-2 large	GPT-2 small Gro-NLP	Llama 13B	Llama 13B chat	GEITje 7B	GEITje 7B chat	GEITje ultra SFT	GEITje ultra DPO	BERTje	RobBERT	RobBERT 2023 large	XLM-RoBERTa	Average
	Average	0.94	0.86	0.91	0.85	0.94	0.92	0.93	0.93	0.89	0.54	0.73	0.82	
Noun Constituent	1. Definite, Indefinite, Categorical and Generic Noun Constituents	0.90	0.84	0.98	0.64	0.94	0.86	0.88	0.90	0.82	0.76	0.72	0.60	0.82
	2. Nominalizations	0.96	0.88	0.96	0.92	1.00	0.96	0.94	0.94	0.88	0.54	0.90	0.90	0.90
Adjectival Constituent	3. The Construction of the Adjective Constituent	0.96	0.82	0.98	0.98	1.00	0.96	1.00	1.00	0.92	0.60	0.58	0.86	0.89
	4. Degree-Indicating or Reinforcing (Pre)Determinations	1.00	0.96	0.94	0.96	0.98	0.98	0.98	0.98	0.96	0.42	0.60	0.96	0.89
	5. Subordinate Clauses	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.99
Adpositional Constituent	6. Pronominal Adverbs	0.88	0.84	0.94	0.98	0.96	0.98	0.96	0.96	0.96	0.42	0.76	0.86	0.88
Verb Constituent	7. Group-Forming and Non-Group-Forming Uses of Verbs	0.98	0.98	0.96	0.96	1.00	1.00	1.00	1.00	0.94	0.24	0.84	0.80	0.89
	8. Verbs With a Participle as a Complement	0.90	0.82	0.78	0.60	0.82	0.80	0.86	0.84	0.68	0.40	0.66	0.68	0.74
	9. Verbs With an Infinitive as a Complement	0.78	0.74	0.66	0.62	0.72	0.68	0.70	0.72	0.74	0.52	0.56	0.66	0.68
	10. The Order Within the Verb Ending Group	1.00	0.86	0.92	0.88	0.98	0.94	0.96	0.96	0.98	0.50	0.66	0.92	0.88

Figure 4: *Accuracy of LLMs on our dataset. We report results for 8 causal language models (first 8 models) and 4 masked language models.*

Overall, transformer-decoder models seemingly perform better on the tested syntactic phenomena than transformer-encoder models. The best performing models were GPT-2 large and GEITje: they both scored an overall accuracy across all syntactic phenomena of 0.94. There are some differences in the performance of these two models across the phenomena, although fairly minimal. Taking a closer look at the syntactic phenomena and the CONSTITUENT types, we see that GPT-2 outperforms GEITje on 3 of the 4 syntactic phenomena belonging to the verbal constituent with the exception of the phenomenon Group-forming and Non-group-forming Uses of Verbs (7) where GEITje had a perfect score and GPT-2 a near-perfect score with 0.98. GEITje, on the other hand, performed better overall when looking at the rest of the constituent types. We can see that GEITje specifically seems to handle nominal constituents better, having scored higher on both syntactic phenomena belonging to this constituent (Definite, Indefinite, Categorical and Generic Nominal Constituents (1) and Nominalizations (2)).

Among the transformer-encoder models, our results show that BERTje has the highest accuracy among the transformer-encoder (masked) LLMs: this is in line with the high model performances reported by Suijkerbuijk et al. (2024), who found BERTje to be the second best-performing model. The dismal performance of RobBERT in our dataset is surprising, given that the model had 0.90 accuracy on BLiMP-NL, and its performance is not on par with the other models (although all the models of the RobBERTa family perform the worst). RobBERT performs perfectly well on subordinate clauses (as most models), while exhibiting moderate to low performance on the other phenomena. While the difference in performance with BLiMP-NL may be largely attributed to the different syntactic phenomena covered in both datasets, we must note that on the overlapping case of Nominalizations, which is present also in BLiMP-NL, the model performed much worse on our dataset – and so did XLM-RoBERTa. Other models, however, did show high accuracy on this phenomenon —and so did human participants— hence it seems fair to expect this may be a result of

the architecture rather than an artifact on our dataset. RobBERT was also found to underperform on semantic similarity by Brans and Bloem (2024).

Furthermore, when looking at the syntactic phenomena, we can see that the LLMs performed best overall on the syntactic phenomenon that covered Subordinate Clauses (5), as part of the Adjective Constituent, with an average accuracy of 0.99. This is a phenomenon that our human participants also had clear opinions about. The average accuracy of the LLMs on the other two syntactic phenomena belonging to the Adjective Constituent were also high, both with a score of 0.89. The LLMs performed worst on the syntactic phenomenon that covered Verbs With an Infinitive as a Complement (9), followed closely by Verbs With a Participle as a Complement (8).

Verbs With an Infinitive as a Complement (9) shows the worst performance overall. These are phenomena for which there is a lot of optionality in Dutch, and human participants also varied in their judgements on this phenomenon. An example of a minimal pair for which most LLMs assigned a higher probability to the wrong sentence in the minimal pair is the following:

(18) Die sportleraar heeft me leren roeien.
That PE-teacher has me taught row.

(19) * Die sportleraar heeft me geleerd te roeien.
That PE-teacher has me taught to row.
'That PE-teacher has taught me to row.'

Of the 12 models, eight assigned a higher probability to the sentence labeled as ungrammatical; these include GPT-2 (-0.32), GPT Large (-0.42), Llama2 (-0.31), Llama2 chat (-0.56), GEITje chat (-0.02), RobBERT (-0.26), RobBERT large (-0.18), and XLM-RoBERTa (-1.31).

Phenomenon 8, covering Verbs With a Participle as a Complement, showed a relatively poor performance for most models. One minimal pair that was incorrectly scored by all LLMs is the following:

(20) De kinderen leren op school zwemmen.
The children learn in school swim.

(21) * De kinderen hebben op school geleerd zwemmen.
The children have in school learned swim.
'The children have learned to swim'.

This case is surprising, as the two sentences differ by several words. However, the ungrammatical sentence is a sentence that is only ungrammatical due to a notoriously complex aspect of Dutch grammar, the *infinitivus pro participio* effect (Augustinus and Van Eynde 2012), where participial verbs are replaced by infinitival ones under certain conditions. The ungrammatical sentence in this pair incorrectly did not undergo IPP. While most of these models showed a relatively small difference between the probabilities they assigned to the sentences in the minimal pair (ranging from -0.08 (by BERTje and GEITje chat) to -0.29 (by GPT2), other models, such as GPT large and XLM RoBERTa, showed a larger difference with a score of -0.71 and -0.61 respectively. The worst performing model for this specific minimal pair was RobBERT large, which showed a difference score of -1.34.

There are some examples that received very clear ratings from human participants, but that models struggled with. These are perhaps the most interesting examples to look at. The aforementioned example 10, which we highlighted in the previous section as one that human participants strongly agree about, was classified incorrectly by six models, including GPT-2 (-0.18), GEITje (-0.1) and Llama2-chat (-1.59). This is another sentence that incorrectly does not have IPP.

The following pair was wrongly scored by many models, and in particular, the normally high-performing GPT-2 model scored it wrong by -0.44:

(22) Nergens stond een titel of datum op.
Nothing stood a title or date on.

(23) * Nergens op stond een titel of datum.
Nothing on stood a title or date.
'Nothing was labeled with a title or date'.

This ungrammatical example concerns an exception that only applies to indefinite adverbs - other types of adverbs can be used in this way. This result might indicate that the models struggle to represent all relevant semantic features in rarer contexts. These results may help to pinpoint linguistic features that models have difficulty representing.

This example was not in our human-rated set but a similar pair was rated fairly well by participants - a convincing 1.38 for the ungrammatical example, though with a less convincing 5.45 for the grammatical counterpart, perhaps indicating that this is a somewhat unusual construction.

5.4 Comparing Performance of Language Models and Native Speakers

Our analysis of the Dutch speaker data revealed that our dataset is generally consistent with the notion of grammaticality of native Dutch speakers, yet some variance is present —particularly for the ungrammatical sentences. We saw that the most interesting cases are the ones where humans and models strongly differ in their judgements. Thus we ask whether humans and models quantitatively correlate in how clearly they distinguish grammatical from ungrammatical sentences in a pair. To address this question, we computed the mean difference between the scores for each grammatical and ungrammatical pair, across participants, and then computed the correlation between the mean and the difference between model probabilities for the same sentences. We perform this analysis for each model and phenomenon, which results in 120 Pearson's r correlation values, shown in Figure 5.

In terms of phenomena, we can see that 9 (Verbs With an Infinitive as a Complement) and 7 (Group-Forming and Non-Group-Forming Uses of Verbs) are the phenomena for which most models exhibit high correlation with human data, although this is reduced to moderate-to-low correlation in the case of transformer-encoder models (with the exception of XLM-Roberta in Verbs with an infinitive as a complement). Interestingly, while most models exhibited a low performance for phenomenon 9, the opposite was the case for phenomenon 7 (again, with the exception of the RobBERT-* models). Yet, both of these phenomena are ones that encompass a lot of optionality. It seems then that the high correlation is not just based on overall performance but likely on various properties of the constituent type, such as its frequency, the extent to which it allows optionality and the type and number of features that grammatical patterns in the constituent type depend on.

On the other hand, the phenomenon (2) Nominalizations was among the phenomena with the lowest acceptability ratings by our Dutch participants. Interestingly, the low frequency of this construction does not seem to pose a problem for LLMs (with the exception of RobBERT, which has a low overall performance). The correlations for this phenomenon are mostly moderate-to-low, reflecting this mismatch. It appears that, at least for this case, LLMs capture prescriptive low-frequency constructions. This could be attributed to various factors: the localized nature of nominalization in the phrase structure, transfer learning from other languages that use more nominalization, or the stylistic nature of the training data – formal writing generally uses more nominalizations and LLMs would encounter this more than most humans. Such hypotheses would require follow-up experiments to confirm.

6. Discussion

Overall, we see that transformer-decoder Dutch LLMs perform better than transformer-encoder models when evaluated on our set of minimal pairs. The emergence of GPT-2 as a winner (on

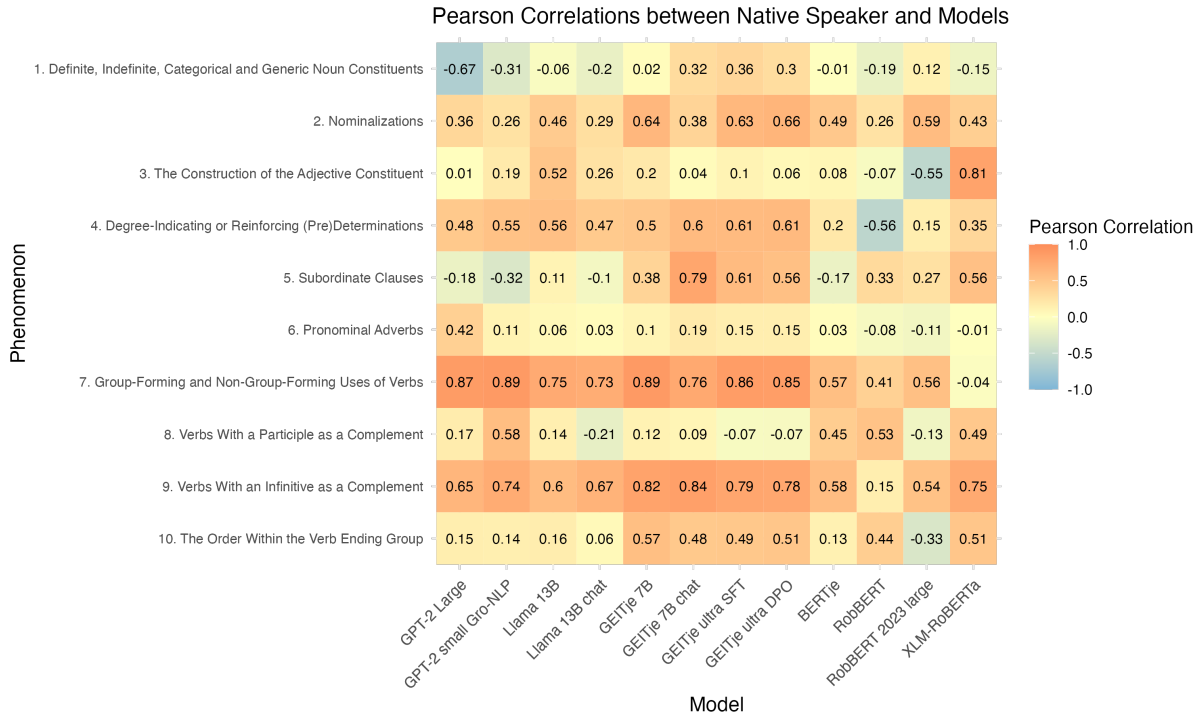


Figure 5: *Pearson's* r *correlation between the native speakers difference scores for sentences in a pair, and the models' difference probability for sentence in a pair.*

par with GEITje) is in line with prior work (Warstadt et al. (2020), Someya and Oseki (2023), Suijkerbuijk et al. (2024)). This is relevant insofar all the LLM-generated minimal pairs used in BLiMP-NL have been generated by a model of the GPT family, which may have given an advantage to GPT-2; in our study, this is to some extent mitigated thanks to the inclusion of ANS minimal pairs and our following of ANS criteria when generating additional pairs.

Our study also breaks the tie between transformer-decoder GPT-2 and transformer-encoder BERTje reported in BLiMP-NL. Nevertheless, BERTje still exhibits a 89% accuracy on our dataset, which is below the 94% of GPT-2 but superior to GPT-2 small and Llama 13B chat, the least performing transformer-decoder models. Other transformer-encoder models, particulary those of the RobBERT family, performed the worst on our dataset, in contrast to the performance of this model in BLiMP-NL. Unsurprisingly, we observe that RobBERT models trained on larger dataset perform better than those trained on smaller ones, yet even the largest of them (XLM-RoBERTa) is less competitive than the rest of the models we evaluate. This is particularly noticeable when dealing with phenomena 1 (definite, indefinite, categorical and generic nominal constituents), 8 (verbs with a participle as a complement), and 9 (verbs with an infinitive as a complement). As discussed before, this is likely due to the very specific selection of grammatical contrasts in Verbal Constituents targeted in the ANS, which centres around optional elements.

Suijkerbuijk et al. (2024) deemed GPT-2 as highly correlating to human grammaticality judgements. In our work, we extend the comparison with human data to include all the models from our evaluation. Interestingly, we found that in fact other models have more correlation with the data from the Dutch speakers, on our dataset — GPT-2 is one of the models with the least correlation to human responses. Note that the overall accuracy is still higher for humans (which is in fact 100%, since the ratings for grammatical sentences are higher than the ratings for ungrammatical sentences

in every pair). The mismatch stems from the *degree* of grammaticality attributed to each sentence (or, more specifically, the difference between the degree of grammaticality for each sentence in a pair). Thus the pairs for which this difference is larger (and therefore the distinction more clear) for models are often not the same pairs with the clearest grammatical distinctions for humans. However, there is a slight difference in that, for humans, the sentences were distributed across different participants, while in the case of models, one single instance of each model type rated all the pairs.

One of the reasons we opted for the use of a dataset of standard Dutch was precisely to find out whether LLMs have learnt the fixed grammatical rules in standard Dutch which, while known to the Dutch-speaking population, occasionally deviates from actual usage and, as we have found, also from human acceptability judgements. Based on the observations depicted above, a relevant question is whether the models which have shown a more human-like linguistic behaviour on our dataset would turn out to be the same on BLiMP-NL, which has been created with semi-automatically generated sentences; however, this analysis remains to be pursued.

7. Conclusion

In this study we have introduced a novel challenge set of linguistic minimal pairs for Dutch, aimed at the evaluation of Dutch grammatical abilities of LLMs. Our challenge set is based on the ANS, hence we provide a benchmark of minimal pairs that has its roots in standard documentation of Dutch grammar, crafted by linguists. We evaluated the performance of multiple transformer-encoder and transformer-decoder models by calculating their accuracy when assigning probabilities to the grammatical and ungrammatical sentences in each minimal pair of the challenge set. Hence, we are able to determine which models behave according to the rules of Dutch grammars that are covered by our dataset.

Moreover, we collected grammatical acceptability judgements from Dutch native speakers on a subset of our challenge set. Participants largely identified the grammatical contrasts captured in our datasets, but also exhibited a small deviation from idealized scores which was more prominent in some syntactic phenomena than others. This variance was instrumental in determining which LLMs exhibited similar sensitivities to humans, at the sentence level. Notably, we found that best performing models do not necessarily correlate with human acceptability ratings, particularly when prescriptive grammar and frequency of use do not align (as is the case of Nominalizations in Dutch); hence in such cases, models may not be representatives of native speakers' intuitions of Dutch grammar.

Acknowledgements

We are grateful to the University of Amsterdam's BSc programme in Cognition, Language and Communication for funding the presentation of this paper, and to the anonymous reviewers for their insightful comments and suggestions.

References

Augustinus, Liesbeth and Frank Van Eynde (2012), A treebank-based investigation of IPP-triggering verbs in Dutch, *Proceedings of TLT*, Vol. 11, pp. 7–12.

Bacon, Geoffrey I (2020), *Evaluating linguistic knowledge in neural networks*, PhD thesis, University of California, Berkeley.

Bakarov, Amir (2018), A survey of word embeddings evaluation methods, *arXiv preprint arXiv:1801.09536*.

Barbiers, Sjef, Hans Bennis, and Lotte Dros-Hendriks (2018), Merging verb cluster variation, *Linguistic Variation* **18** (1), pp. 144–196, John Benjamins Publishing Company Amsterdam/Philadelphia.

Belinkov, Yonatan and James Glass (2019), Analysis methods in neural language processing: A survey, *Transactions of the Association for Computational Linguistics* **7**, pp. 49–72, MIT Press.

Bin-Nashwan, Saeed Awadh, Mouad Sadallah, and Mohamed Bouteraa (2023), Use of ChatGPT in academia: Academic integrity hangs in the balance, *Technology in Society* **75**, pp. 102370. https://www.sciencedirect.com/science/article/pii/S0160791X23001756.

Bloem, Jelke, Arjen Versloot, and Fred Weerman (2017), Verbal cluster order and processing complexity, *Language Sciences* **60**, pp. 94–119, Elsevier.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy (2003), *Probabilistic linguistics*, MIT press.

Brans, Lizzy and Jelke Bloem (2024), SimLex-999 for Dutch, *in* Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, pp. 14832–14845. https://aclanthology.org/2024.lrec-main.1292.

Broekhuis, Hans (2012), *Syntax of Dutch: Nouns and Noun Phrases, Volume 2 (Volume 1.0)*, Vol. 1 of *Comprehensive Grammar Resources*, 1st ed. ed., Amsterdam University Press, Netherlands.

Bylinina, Lisa and Alexey Tikhonov (2022), The driving forces of polarity-sensitivity: Experiments with multilingual pre-trained neural language models, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

Chomsky, N (1965), Aspects of the theory of syntax, *Special technical report. Research laboratory of electronics. Massachusetts institute of technology*, Cambridge: MIT.

Colleman, T., J. De Caluwe, W. Haeseryn, R. Boogaart, F. Landsbergen, and J. Van Hoorde (2021), Over de Algemene Nederlandse Spraakkunst. Accessed: 2023-12-17. https://e-ans.ivdnt.org/over.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020), Unsupervised cross-lingual representation learning at scale, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451.

Corver, Norbert, Hans Broekhuis, and Riet Vos (2015), *Syntax of Dutch: Verbs and Verb Phrases, Volume 1*, Comprehensive Grammar Resources, University Press, Amsterdam.

Coussé, Evie, Mona Arfs, and Gert De Sutter (2008), Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep: een taalgebruiksgebaseerd perspectief op de synchronie en diachronie van de zgn. rode en groene woordvolgorde, *in* Rawoens, Gudrun, editor, *Taal aan den lijve: het gebruik van corpora in taalkundig onderzoek en taalonderwijs*, Academia Press, pp. 29–47.

de Vries, Wietse and Malvina Nissim (2021), As good as new. How to successfully recycle English GPT-2 to make models for other languages, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, p. 836–846. http://dx.doi.org/10.18653/v1/2021.findings-acl.74.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), BERTje: A Dutch BERT Model, *Technical Report arXiv:1912.09582*, arXiv. arXiv:1912.09582 [cs] type: article. http://arxiv.org/abs/1912.09582.

de Vries, Wietse, Martijn Wieling, and Malvina Nissim (2023), DUMB: A Benchmark for Smart Evaluation of Dutch Models, *in* Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 7221–7241. https://aclanthology.org/2023.emnlp-main.447.

Delobelle, Pieter and François Remy (2024), RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion, *Computational Linguistics in the Netherlands Journal* **13**, pp. 193–203. https://www.clinjournal.org/clinj/article/view/180.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: a Dutch RoBERTa-based language model, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3255–3265.

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)*, pp. 4171–4186.

Divjak, Dagmar (2017), The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in Polish, *Cognitive science* **41** (2), pp. 354–382, Wiley Online Library.

Featherston, Sam (2005), The Decathlon model of empirical syntax, *Linguistic Evidence* pp. 187–208, Mouton de Gruyter.

Francis, Elaine (2022), *Gradient acceptability and linguistic theory*, Vol. 11 of *Oxford Surveys in Syntax and Morphology*, Oxford University Press.

Goodall, Grant (2021), *The Cambridge handbook of experimental syntax*, Cambridge University Press.

Haeseryn, Walter, Kirsten Romijn, and Guido Geerts (1984), *Algemene Nederlandse Spraakkunst*, Wolters-Noordhoff.

Havinga, Yeb (2024), GPT2-Large pre-trained on cleaned Dutch mC4. https://huggingface.co/yhavinga/gpt2-large-dutch.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015), SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation, *Computational Linguistics* **41** (4), pp. 665–695. https://doi.org/10.1162/COLI_a_00237.

Hochreiter, Sepp and Jürgen Schmidhuber (1997), Long short-term memory, *Neural computation* **9**, pp. 1735–80.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed (2023), Mistral 7B. arXiv:2310.06825 [cs]. http://arxiv.org/abs/2310.06825.

Kalyan, Katikapalli Subramanyam (2024), A survey of GPT-3 family large language models including ChatGPT and GPT-4, *Natural Language Processing Journal* **6**, pp. 100048, Elsevier.

Kauf, Carina and Anna Ivanova (2023), A better way to do masked language model scoring, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 925–935.

Kruit, Benno (2023), Minimalist entity disambiguation for mid-resource languages, *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 299–306.

Kudugunta, Sneha, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat (2023), Madlad-400: A multilingual and document-level large audited dataset, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 67284–67296.

Lau, Jey Han, Alexander Clark, and Shalom Lappin (2017), Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge, *Cognitive science* **41** (5), pp. 1202–1241, Wiley Online Library.

Likert, R. (1932), A technique for the measurement of attitudes, *Archives of Psychology* **22 140**, pp. 55–55.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. http://arxiv.org/abs/1907.11692.

Micheletti, Nicolo, Samuel Belkadi, Lifeng Han, and Goran Nenadic (2024), Exploration of Masked and Causal Language Modelling for Text Generation. arXiv:2405.12630 [cs]. http://arxiv.org/abs/2405.12630.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth (2023), Recent advances in natural language processing via large pre-trained language models: A survey, *ACM Computing Surveys* **56** (2), pp. 1–40, ACM New York, NY.

Nyandwi, Jean (2023), The Transformer Blueprint: A holistic guide to the transformer neural network architecture, *Deep Learning Revision*. https://deeprevision.github.io/posts/001-transformer/.

Olthof, Marieke, Maud Westendorp, Jelke Bloem, and Fred Weerman (2017), Synchronic variation and diachronic change in Dutch two-verb clusters, *Tijdschrift voor Nederlandse Taal- en Letterkunde* **133** (1), pp. 34–60, Uitgeverij Verloren.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko

Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph (2024), GPT-4 Technical Report. arXiv:2303.08774 [cs]. http://arxiv.org/abs/2303.08774.

Papadimitriou, Isabel, Kezia Lopez, and Dan Jurafsky (2023), Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models, *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1194–1200.

Qualtrics (2005), Qualtrics. Accessed: June 2024.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019), Language models are unsupervised multitask learners, *OpenAI blog*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020), Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of machine learning research* **21** (140), pp. 1–67.

Rijgersberg, Edwin (2023), GEITje: een groot open Nederlands taalmodel. https://github.com/Rijgersberg/GEITje.

Schwartz, Ilan S, Katherine E Link, Roxana Daneshjou, and Nicolás Cortés-Penfield (2024), Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation, *Clinical Infectious Diseases* **78** (4), pp. 860–866. https://doi.org/10.1093/cid/ciad633.

Someya, Taiga and Yohei Oseki (2023), JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs, *in* Vlachos, Andreas and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, Association for Computational Linguistics, Dubrovnik, Croatia, pp. 1581–1594. https://aclanthology.org/2023.findings-eacl.117.

Song, Yixiao, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer (2022), SLING: Sino Linguistic Evaluation of Large Language Models, *in* Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 4606–4634. https://aclanthology.org/2022.emnlp-main.305.

Sprouse, Jon (2023), *The Oxford handbook of experimental syntax*, Oxford University Press.

Suijkerbuijk, Michelle, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank (2024), BLiMP-NL: A corpus of Dutch minimal pairs and grammaticality judgements for language model evaluation, *PsyArXiv*. https://europepmc.org/article/ppr/ppr837942.

Swerts, Marc and Carel van Wijk (2005), Prosodic, lexico-syntactic and regional influences on word order in Dutch verbal endgroups, *Journal of Phonetics* **33** (2), pp. 243–262, Elsevier.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023), Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288.

Vanroy, Bram (2023), Language resources for Dutch large language modelling, *arXiv preprint arXiv:2312.12852*.

Vanroy, Bram (2024), GEITje 7B Ultra: A conversational model for Dutch. https://arxiv.org/abs/2412.04092.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017), Attention is All you Need, *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc.

Veenboer, Tim and Jelke Bloem (2023), Using collostructional analysis to evaluate BERT's representation of linguistic constructions, *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12937–12951.

Vlantis, Daniel and Jelke Bloem (2025), Intrinsic evaluation of mono- and multilingual Dutch language models, *Computational Linguistics in the Netherlands Journal*.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman (2020), BLiMP: The benchmark of linguistic minimal pairs for English, *Transactions of the Association for Computational Linguistics* **8**, pp. 377–392, MIT Press.

Xiang, Beilei, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann (2021), CLiMP: A Benchmark for Chinese Language Model Evaluation, *in* Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, Online, pp. 2784–2790. https://aclanthology.org/2021.eacl-main.242.

Zhu, Wenhao, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li (2023), Extrapolating large language models to Non-English by aligning languages. https://arxiv.org/abs/2308.04948.

Appendix A. Summary of Human Participants Responses

Table 3 reports descriptive statistics of the behavioural responses in our experiment with Dutch native speakers.

Ph.	Gr.	Mean Average	Standard Deviation	n	Standard Error
1	0	1.600427	0.404882	9	0.134961
1	1	5.483226	1.360018	18	0.320559
2	0	2.044444	0.780331	9	0.260110
2	1	5.706303	0.990645	18	0.233497
3	0	2.014646	0.425550	9	0.141850
3	1	6.040783	0.702612	18	0.165607
4	0	2.580051	0.730246	9	0.243415
4	1	4.850505	1.449643	18	0.341684
5	0	2.083333	1.207334	9	0.402445
5	1	4.895833	0.938015	18	0.221092
6	0	3.211111	1.146606	9	0.382202
6	1	6.016667	1.115862	18	0.263011
7	0	1.965629	0.475893	9	0.158631
7	1	5.968575	0.874535	18	0.206130
8	0	2.658069	1.591188	9	0.530396
8	1	5.179894	1.545798	18	0.364348
9	0	2.158730	1.102987	9	0.367662
9	1	5.581349	0.732688	18	0.172696
10	0	2.438412	1.289489	9	0.429830
10	1	5.469066	1.029090	18	0.242559

Table 3: *Summary of the native speaker statistics across all syntactic phenomena. The phenomena (column 'Ph.') are numbered 1 through 10, with the full names of the phenomena as follows: 1. Definite, Indefinite, Categorical and Generic Nominal Constituents, 2. Nominalizations, 3. The Construction of the Adjective Constituent, 4. Degree-Indicating or Reinforcing (Pre)Determiner, 5. Subordinate Clauses, 6. Pronominal Adverbs, 7. Group-Forming and Non-Group-Forming Uses of Verbs, 8. Verbs With a Participle as a Complement, 9. Verbs With an Infinitive as a Complement, 10. The Order Within the Verb Ending Group. A value of 1 in the "Gr." column indicates grammatical, and 0 indicates ungrammatical. For each phenomenon, we provide the Mean Average, Standard Deviation, and Standard Error of the acceptability ratings per grammaticality. n indicates the number of sentences per category.*

Appendix B. Dataset Extension

Table 4 shows the sentences of the ANS which required slight modifications to ensure an accurate reflection of the syntactic phenomenon.

Phenomenon	Correct sentence (ANS)	Wrong sentence (ANS)	Correct sentence (Revised)	Wrong sentence (Revised)
3	(een) - hard (gesteente)	(een) zeer - (gesteente)	Een zeer hard gesteente.	Een zeer gesteente.
3	(Hij is bepaald niet) tuk op vergaderen.	(Hij is bepaald niet) tuk -	Hij is niet bepaald tuk op vergaderen.	Hij is niet bepaald tuk.
3	(Hij is bepaald niet) tuk op vergaderen.	(Hij is bepaald niet) - op vergaderen.	Hij is niet bepaald tuk op vergaderen.	Hij is niet bepaald op vergaderen.
4	(Dat is) makkelijk zat.	(een) makkelijk zat (klusje)	Dat klusje is makkelijk zat.	Een makkelijk zat klusje.
4	NA	(Ik vind het een) hele handig (apparaat.	Ik vind het een heel handig apparaat	Ik vind het een hele handig apparaat.
4	NA	(Die flats zijn) geweldige hoog.	Die flats zijn geweldig hoog.	Die flats zijn geweldige hoog.
6	Waarmee kan ik u helpen?	[E]rgens in moet je tenslotte geloven.	Je moet tenslotte ergens in geloven.	Ergens in moet je tenslotte geloven.
6	Waarmee kan ik u helpen?	Nergens op wordt echt diep ingegaan.	Nergens wordt echt diep op ingegaan.	Nergens op wordt echt diep ingegaan.
6	Waarmee kan ik u helpen?	Overal over heeft hij een mening.	Hij heeft overal een mening over.	Overal over heeft hij een mening.
6	Het gebouw én het huis ernaast verdwijnen en maken plaats voor luxeappartementen.	het huis naast het	Het gebouw én het huis ernaast verdwijnen en maken plaats voor luxeappartementen.	Het gebouw en het huis naast het verdwijnen en maken plaats voor luxeappartementen.
6	Ria Dekker . . . vist meestal in oude kleren, met een regenbroek erover.	met een regenbroek over ze	Ria Dekker draagt meestal in oude kleren, met een regenbroek erover.	Ria Dekker draagt meestal in oude kleren, met een regenbroek over ze.
9	NA	Ze hebben over het hekje te praten gehangen.	Ze hebben over het hekje hangen te praten.	Ze hebben over het hekje te praten gehangen.
10	Ze zag dat er een politieauto kwam aangereden.	NA	Ze zag dat er een politieauto kwam aangereden.	Ze zag dat er een politieauto aangereden kwam.
10	Ik hoop dat we voor zo'n ramp gespaard blijven.	NA	Ik hoop dat we voor zo'n ramp gespaard blijven.	Ik hoop dat we voor zo'n ramp blijven gespaard.

Table 4: *Minimal pairs extracted from e-ANS that were modified.*