Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French

Ayla Rigouts Terryn[1], Miryam de Lhoneux[2]
[1]KU Leuven, Centre for Computational Linguistics (CCL)
[2]KU Leuven, Department of Computer Science
ayla.rigoutsterryn@kuleuven.be, miryam.delhoneux@kuleuven.be

Abstract

The most widely used LLMs like GPT4 and Llama 2 are trained on large amounts of data, mostly in English but are still able to deal with non-English languages. This English bias leads to lower performance in other languages, especially low-resource ones. This paper studies the linguistic quality of LLMs in two non-English high-resource languages: Dutch and French, with a focus on the influence of English. We first construct a comparable corpus of text generated by humans versus LLMs (GPT-4, Zephyr, and GEITje) in the news domain. We proceed to annotate linguistic issues in the LLM-generated texts, obtaining high inter-annotator agreement, and analyse these annotated issues. We find a substantial influence of English for all models under all conditions: on average, 16% of all annotations of linguistic errors or peculiarities had a clear link to English. Fine-tuning a LLM to a target language (GEITje is fine-tuned on Dutch) reduces the number of linguistic issues and probably also the influence of English. We further find that using a more elaborate prompt leads to linguistically better results than a concise prompt. Finally, increasing the temperature for one of the models leads to lower linguistic quality but does not alter the influence of English.

Keywords: LLM, bias, cross-lingual

1. Introduction

In recent years, (generative, pre-trained) large language models (LLMs) have substantially advanced and changed the field of natural language processing (NLP), with large models displaying an "unusually large set of capabilities" (Tamkin et al., 2021) across a wide range of tasks, including acting as a chatbot. Their capabilities and ease of use have contributed to a quick rise in popularity, including among non-expert users. For instance, a recent report on the use of digital technologies in Flanders in 2023 (De Marez et al., 2024) showed that 18% of people in this region use a tool to generate text, music, images, or speech at least monthly. For chatbots specifically, this number drops a little to 14%. Given how recently AI chatbots have become available, this illustrates how fast they are gaining influence.

The undeniably impressive capabilities of the LLMs behind AI chatbots do not imply the technology is without its flaws. For instance, the production of false content by LLMs is common enough that it quickly got a dedicated term: *hallucinations* (see, e.g., Ye et al. (2023)). The models are also known to be *biased* (see, e.g., Vig et al. (2020)). A third issue, which constitutes the central theme of this study, pertains to the *English bias*. This refers to the tendency for LLMs to be predominantly trained on English datasets. The problem goes beyond LLMs, and affects NLP in general: "[e]xisting estimates of how much of top venue NLP research is devoted to English vary a bit, but typically lie in the range of 50-90%" (Søgaard, 2022, p.5254).

The English bias has many effects. Logically, the performance of NLP tools is often best in English. This is clearly illustrated for machine translation, where performance tends to be highest for language pairs that include English, for translation into English, and for English in combination with a closely related language, as illustrated by, e.g., the results of WMT23 (Kocmi et al., 2023). However, this English bias goes beyond performance issues. For instance, De Bruyne (2023) argues that the predominance of English has a (negative) impact on the conceptualisation of emotion detection, as emotions and the ways people verbalise emotions are not universal.

An effect that has not been researched extensively is the linguistic quality of texts generated by LLMs in languages other than English and, specifically, whether and how English bias influences these texts (e.g., presence of anglicisms). The latter is a well-known issue among attentive non-English users of the technology, but very little research can be found where the issue is officially established and analysed. The main goal of this exploratory study is to document general linguistic issues in texts written by generative LLMs and to analyse how often these issues might be traced back to the English bias. The secondary goal is to provide a starting point for future (more extensive) research by testing a methodology based on human annotations and starting to identify the role of some of the main variables, such as the models (and their training data and sizes), languages,

Proceedings of the HumEval @LREC-COLING 2024, pages 12–27
Turin, Italy. May 21st, 2024.

temperature, and prompts.

A brief overview of related research can be found in Section 2. The methodology is described in Section 3, with separate subsections on corpus creation and annotation. Section 4 is dedicated to the findings. Limitations, conclusions, and opportunities for future research are discussed in Section 5.

2. Related Research

The most widely used LLMs like GPT4 (OpenAI et al., 2023) and Llama 2 (Touvron et al., 2023) are trained on large amounts of mostly English data, but are still able to deal with non-English languages (Shi et al., 2023). This English bias leads to lower performance in other languages, especially for low-resource languages (e.g. Hendy et al., 2023, among others) and for tasks that are not translatable (Zhang et al., 2023). This has led researchers to speculate that these models use English as a pivot language in which they reason, prior to generating output in non-English languages. Wendler et al. (2024) empirically test this speculation by inspecting internal model representations via mechanistic interpretability. They develop tasks (translation, repetition and a cloze task) where the output is expected to be in non-English languages (here mainly Chinese, but with controlled experiments in French and Russian) and investigate the latent representations of Llama models at the different layers. They find evidence that the representations in the intermediate layers of these models are closer to English than to other languages, confirming that English may act as a pivot language. Contemporaneously, Zhao et al. (2024) probe LLMs for language-specific information leading them to very similar findings.

Our work focusses on the analysis of the model *outputs*. While Wendler et al. (2024) find that the influence of English in the representation declines to a very small percentage in the last layers of the models, we find clear traces of it in the output. This complements the evidence that English is used as a pivot language in these models. We further contribute a characterisation of *how* English manifests itself in model outputs in non-English languages, here Dutch and French.

3. Methodology

3.1. Corpus

With the intended goal of analysing linguistic output of generative LLMs in non-English languages and the impact of English bias, we decided on a corpus-based approach with expert annotations as the best way to obtain nuanced and fine-grained information. Ideally, a comparable corpus of human-written and LLM-generated texts would allow for a controlled comparison. Because data-driven systems like LLMs perform best on content that is well represented in the training data, we want to work with common text types in the general domain, to avoid adding domain-specific difficulties. At this stage, well-resourced languages that are closely related to English (so, probably some of the *easiest* languages besides English for these models to handle) were preferred. First, because the goal is to include multiple models, we are limited by the availability of models that include the languages. Second, because the quality of texts written by LLMs in low-resource languages can be too low to allow a detailed analyses. Third, because it is worth seeing which problems remain even in such ideal settings. Ideally, lessons learnt from this exploratory study can be used to launch similar studies for low(er) resourced languages.

The languages of choice based on these criteria were Dutch and French, specifically the dialect clusters from Belgium: Flemish and Belgian French. Both are well established and standardised national languages with very limited variation from Dutch and French in the Netherlands and France. An advantage is that it allows a selection of equivalent sources in the same country for both languages. The text collection was based on insights from Schepens et al. (2023) and Muñoz-Ortiz et al. (2023). The former create a German corpus of LLM-generated texts based on prompts including children's books titles, a request to write in German, and for children within a certain age range. The latter use (English) newspaper headlines and the first three words of the article as prompts. Newspaper headlines fit the criteria of general text types that should be well represented in the training data of most models. To create a comparable corpus of publicly available texts in Dutch and French, the online newspapers of the Belgian public broadcasting companies were selected as sources, with articles from the Flemish (Dutch) *Vlaamse Radio- en Televisieomroeporganisatie (VRT)* and from its (Belgian) French counterpart, *Radio-Télévision belge de la Communauté française (RTBF)*.

In terms of the choice of models and settings, there are many potentially influential variables (model architecture, model size, training data size, language distribution in training data, temperature, top P, prompt, etc.). In a small-scale study like this, it is impossible to control for everything and to obtain enough useful data with all potentially relevant combinations of settings. The practical limitations (time and budget) allowed about 500 articles to be annotated. With 50 articles per setting as a reasonable estimate for the minimum volume of text

	Prompt A	Prompt B
Dutch	Je bent een ervaren journalist bij VRT NWS, de nieuwssite van de Vlaamse openbare omroep. Je moedertaal is Nederlands (Vlaams). Schrijf een artikel voor VRT NWS op basis van volgende titel: [title]	Schrijf een artikel op basis van volgende titel: [title]
French	Tu travailles en tant que journaliste pour la RTBF, la référence francophone de l'actualité publique belge, et tu as beaucoup d'expérience. Ta langue maternelle est le français (de Belgique). Ecris un article pour la RTBF ayant le titre suivant : [title]	Ecris un article ayant le titre suivant: [title]
English equivalent	You are an experienced journalist working for [name of broadcasting company], the news website of the [Flemish or Belgian French] public broadcaster. Your native language is [Dutch (Flemish) or French (from Belgium)]. Write an article for [name of broadcaster] based on the following title: [title]	Write an article based on the following title: [title]

Table 1: Elaborate (A) and concise (B) prompts used in Dutch and French, incl. English translation

required for a meaningful analysis, this amounted to 10 different experimental settings. 50 articles were collected in Dutch and French respectively, spread over various categories of news (national, international, sports, politics, etc.) and making sure the subjects were equivalent in both languages. An overview of the original articles and sources has been added in the appendix. With few exceptions (to find equivalent articles in Dutch and French), only recently published articles were selected to limit the chances of them being included in the training data of the models.

Though we cannot control for all differences between available pretrained models, in the context of this project we looked for (1) one of the largest, best performing models as a reflection of what is currently possible, (2) one (smaller) open source model that allows further research, and (3) one model with more fine-tuning on the non-English language to see whether and how much this can improve results. As *prompt engineering* has also been shown to be influential (White et al., 2023), two different prompts were chosen as additional variables: one elaborate prompt that considers common insights from prompt engineering, like assigning a role (prompt A), and one very concise prompt (prompt B). The exact prompts and an English translation can be found in Table 1. However, as this doubled the number of experiments, to limit the number of articles to 500, the decision was made to only include a fine-tuned (language-specific) model for Dutch, as the lesser-resourced of the two languages. This means the project includes 3 models, all of which are used for Dutch, and two of which are used for French:

- **GPT-4** (OpenAI et al., 2023):
  - Settings: used in OpenAI Playground (chat), temperature=1.0, maximum_length=8000, top_P=1.
  - Motivation: one of the most powerful and influential models available at the time of the experiment (Zhao et al., 2023).
  - Limitations: not open source.
- **Zephyr 7B Beta** (Tunstall et al., 2023):
  - Settings: used in the HuggingFace chat version[1], temperature=0.7, max_new_tokens=1024 (+ click *continue generating* when option is provided after incomplete response), top_P=0.95.
  - Motivation: One of the best-performing open source models for Dutch based on (Vanroy, 2023), without specific fine-tuning for Dutch (based on Mistral (Jiang et al., 2023)).
  - Limitations: trained on synthetic datasets and more likely to generate problematic content according to the technical report, despite high scores on truthfulness tasks (Vanroy, 2023).
- **GEITje Chat V2 7B** (Rijgersberg and Lucassen, 2023) (only for Dutch):
  - Settings: used in LM Studio[2], temperature=2.0, n_predict=-1 ("to allow the model to stop on its own"), top_P=0.95.
  - Motivation: open source model specifically fine-tuned for Dutch (also based on Mistral).
  - Limitations: no preference optimisation and small for a LLM; GEITje-7B-ultra is superior as a chatbot, but was published after experiments had already started.

[1] https://huggingface.co/spaces/HuggingFaceH4/zephyr-chat

[2] https://lmstudio.ai/

source of articles				av. #		
model	**tmp**	**l**	**p**	**tok**	**typ**	**typ/tok**
GEITje	0.2	NL	A	170	77	0.59
			B	127	66	0.67
	0.85		A	136	82	0.68
GPT-4	1.0	FR	A	449	233	0.52
			B	450	217	0.48
		NL	A	394	198	0.50
			B	440	212	0.48
Zephyr	0.7	FR	A	560	320	0.59
			B	594	334	0.58
		NL	A	494	276	0.59
			B	528	311	0.59
VRT (Dutch)				494	217	0.47
RTBF (French)				441	202	0.50

Table 2: Average (av.) number of tokens (tok), types (typ) (lowercased), and type/token ratio per part of the corpus, distinguishing between model, temperature (tmp), language (l), and prompt (p)

For each model, the default (recommended) settings were selected, except for the maximum length, which was set to the maximum allowed value, so the systems were able to write articles of lengths comparable to those of the original articles. All texts were generated between the 2nd and 31st of January 2024. Because the recommended temperature for GEITje is so much lower than for the other models, some experiments were duplicated using the same settings but a higher temperature (0.85, which is between 1.0 (for GPT-4) and 0.7 (Zephyr)). The result is a collection of 550 articles generated by the LLMs, based on the titles of 50 Dutch and 50 French articles written by human journalists. GEITje had to be stopped manually five times because the systems appeared to be stuck endlessly generating the (exact) same paragraphs. The overview, along with token counts, to indicate the size of the corpus can be found in Table 2. A discussion of these numbers and the type/token ratio can be found in Section 4.

3.2. Annotation

3.2.1. Annotation scheme

As mentioned, the goal of this project is to establish and document linguistic peculiarities (both clear errors and any text that could be seen as problematic from a linguistic perspective), and to analyse how often issues might be traced back to English. Based on preliminary observations by the leading researcher, an annotation scheme was established to divide these observations into nine categories with labels to allow a nuanced analysis:

- English word/phrase
  - not usually used in Dutch/French
  - sometimes used in Dutch/French
  - very commonly used in Dutch/French
- longer piece of English text
  - part of text
  - entire text
- word/phrase does not exist (*)
- grammar mistake (*)
- spelling mistake (*)
- strange/wrong construction (*)
- strangely used word/phrase (*)
- other linguistic remark
- non-linguistic remark

Options marked with (*) all have three labels:
  - clearly from English
  - could be from English
  - no clear link to English

There are 2 additional markers: '*Not sure*' and '*Very minor mistake/humans might write the same*'. More detailed information, including examples for each category, can be found in the annotation guidelines.[3] The category for non-linguistic remarks was added to allow annotators to mark strange or nonsensical text passages, even when the issue is not linguistic, but they were instructed to keep this for *meta* information (e.g., the language model writing that it is a language model), or very obviously wrong information that feels weird not to mark (e.g., calling penguins mammals). During the annotation, the annotators did not see the source of the articles, so they could not develop a bias, e.g., when realising that some systems consistently write better or worse texts. All annotations were made in Label Studio (Tkachenko et al., 2020-2022).

3.2.2. Annotators

Professional translators with experience translating from English were hired to perform the annotations in their native languages because: (1) translators are assumed to know both their source and target languages very well, (2) translators are supposed to be especially attentive to influences from their source language into their target language, and (3) translators have experience revising and (post-)editing (translated) texts, which can be seen as relevant experience for this task. There were two main annotators: one who annotated all French texts, and one who annotated all Dutch texts. All annotators are native speakers of either Flemish Dutch or Belgian French.

[3] https://github.com/AylaRT/English_bias_annotation_guidelines.git

3.2.3. Inter-annotator agreement

Besides the main annotators, two additional annotators (one professional translator, one researcher with a background in translation; both native speakers) were included to calculate inter-annotator agreement (IAA). The main Dutch annotator and the extra annotators all annotated the same 21 Dutch articles based on the first three Dutch titles.

The first problem with calculating IAA is the lack of a minimum or maximum number of possible annotations, excluding many commonly used metrics. The second problem is that the span selection was not very rigid, both because it can be difficult and not many guidelines were defined in this respect, and because annotators were not always careful about including or excluding trailing spaces. This means that automatic calculations offered within Label Studio (e.g., *basic matching function*: [4]) are quite pessimistic, with agreement scores between 45% and 50%. Therefore, part of the IAA calculation was done manually, examining all annotations and matching them if they were clearly about the same item, even when spans did not overlap perfectly (e.g., annotation of *worst-case scenario's*, or only *worst-case* as English words in Dutch, because the word *scenario* is the same in both languages). The result is a list of 187 possible annotations, with for each possible annotation and annotator an indication of whether the instance was annotated, and, if so, which category was used with which label(s).

This analysis shows good agreement on whether to annotate: annotator pairs agree for 73% to 83% of all 187 items. All three agree on 67% of the items. As this does not consider all of the times where none of the annotators mark anything, this is good agreement. One annotator (not the main one) annotates slightly more than the other two (170 versus 147 and 148 annotations respectively). Out of 62 annotations for which at least one annotator disagrees, 21 are marked as *minor* or *not sure*.

Annotator pairs also agree on which category to assign for 65% to 79% of all 187 instances. The confusion matrices show relatively good agreement overall, with a few logical patterns. One of the matrices is shown in Table 3. The others can be found in the appendix. One annotator is stricter than the others, e.g., annotating wrong punctuation. The most ambiguous categories are *strangely/wrongly used phrase* and *strange/wrong construction*. This was expected, since annotators cannot easily consult resources like dictionaries or grammars to check whether their instinct that a word, phrase, or construction is strange or wrong, is more than a personal preference. However, even seemingly unambiguous categories like *nonexistent word* and *English word* can be ambiguous, for instance when a Dutch text mentions *gefeed*, i.e., the English word *feed* used with a dutch prefix to conform to Dutch grammar rules. These disagreements are also indications of how the guidelines can be improved in the future, e.g., splitting the rather prescriptive sounding *word/phrase exist* into one category for words/phrases that appear made up by the LLM and have never been written by humans (at least not based on texts that can be found online), and one category for words/phrases that may not be part of the official standard language, but are used by human writers as well.

Most categories include the same 3 labels about potential influence from English, so the agreement on these labels can be compared regardless of the categories. Counting the same labels as perfect agreement, and disagreement with only one point difference as 50% agreement, there was 63% to 77% agreement on the label per annotator pair.

Dutch vs. French main annotators: We can get an idea about agreement for French versus Dutch annotations based on the Dutch IAA analysis. No unexpected differences were found, except for the most ambiguous category *strangely/wrongly used word/phrase*. The French annotator used this category a lot more than the Dutch annotator: 18.4 times/1000 tokens on average, versus only 3.6 times/1000 tokens on average. For *spelling mistake* there is a difference of 4.8, and for all other categories, the difference is below 1.5. This is observed in all settings and only for the most ambiguous category, which leads us to conclude that the French annotator was quicker to annotate *strangely/wrongly used words/phrases*, and that this does not necessarily reflect a difference in performance of the LLMs in French versus Dutch. More research is required to confirm this and to improve comparisons across languages. Thus, cross-lingual comparisons in the current project are limited.

In conclusion, agreement is high enough to use the annotations for an exploratory analysis of the texts, provided known disagreements and ambiguities are carefully considered.

4. Findings

All analyses are based only on the annotations made by the two main annotators (one per language). Since average text length vary per system, the analysis takes this into account and looks at the number of annotations (per category) per 1000 tokens. This works well, except for the Zephyr model in Dutch, especially with the concise prompt (B), because with this setting, Zephyr wrote 36 of the 50 articles completely in English. In those cases, there will only be a single annotation (*entire text in English*). This makes it seem as if there are

[4] https://docs.humansignal.com/guide/stats

annotator A → vs B ↓	English word/phrase	grammar mistake	longer piece of English text	non-linguistic remark	other linguistic remark	spelling mistake	strange/wrong construction	strangely/wrongly used word/phrase	word/phrase does not exist	#NA	Total
English word/phrase	13	0	0	0	0	0	0	0	2	0	**15**
grammar mistake	0	23	0	0	0	0	2	1	0	11	**37**
longer piece of English text	0	0	2	0	0	0	0	0	0	0	**2**
non-linguistic remark	0	0	0	1	0	0	0	0	0	0	**1**
other linguistic remark	0	0	0	4	5	0	0	0	0	4	**13**
spelling mistake	1	0	0	0	0	12	0	0	0	5	**18**
strange/wrong construction	0	0	0	0	0	0	23	0	0	11	**34**
strangely/wrongly used word/phrase	0	0	0	0	0	0	2	26	0	5	**33**
word/phrase does not exist	0	0	0	0	0	1	0	1	14	1	**17**
#NA	1	3	0	0	0	6	1	3		3	**17**
Total	**15**	**26**	**2**	**5**	**5**	**19**	**28**	**31**	**16**	**40**	**187**

Table 3: Confusion matrix based on the annotations of two of the annotators

	GPT-4 temp:1				**GEITje temp:.2**		**GEITje temp:.85**
	FR		**NL**		**NL**		**NL**
av. # annotations per category, per 1000 tokens	**A**	**B**	**A**	**B**	**A**	**B**	**A**
English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34	1.74
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0	0.28
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10	2.86
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26	10.73
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75	4.36
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50	1.70
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45	0.71
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07	4.61
all annotations (excl. non-ling.)	24.08	24.83	15.50	17.10	16.22	19.40	22.37
all annotations	24.97	25.38	16.08	17.98	19.07	20.47	26.98
text written completely in English	0	0	0	1	0	1	0
average % of annotations with:							
clear English influence	7%	6%	8%	13%	24%	6%	4%
potential influence from English	36%	39%	14%	26%	26%	24%	33%
no clear influence from English	57%	54%	78%	60%	50%	70%	63%

Table 4: Averaged findings per setting (language FR or NL; prompt A or B) of GPT-4 and GEITje (with recommended temperature of .2, then with temperature of .85)

very few annotations in the other categories in this setting (because these cannot be annotated in the English texts), which is not representative (the few Dutch texts do contain a lot of annotations). Therefore, this setting is often excluded from the general analyses.

Number of tokens and types: A first observation based on the information in Table 2 is that the average lengths of articles differs substantially. GPT-4's average article length is closest to that of the original articles. GEITje regularly writes articles that consist just of (a rephrasing of) the original title (28 of the 150 articles written by GEITje have <50 words). The type/token ratio is also similar for the original articles and the ones written by GPT-4, but higher for Zephyr and GEITje, indicating those models use a more diverse vocabulary. This is especially noteworthy given the fact that annotators indicated that the generated articles were very repetitive. As mentioned, GEITje was even stopped five times because the systems appeared stuck endlessly generating the exact same paragraphs.

Zephyr: As expected (because it is a smaller model than GPT-4 and not specifically fine-tuned on Dutch like GEITje), the linguistic quality of texts written by Zephyr is clearly the worst out of the three models. As mentioned, it systematically (36 out of 50 prompts) writes an English article when prompted with the concise prompt in Dutch. It does so for the French concise prompt four times as well, and also twice for the Dutch elaborate prompt. Interestingly, this happens less in French than in Dutch, but with the French prompts, there were also two articles written completely in German and one in Spanish. When writing in the expected language, there are still regularly longer pieces of texts written in English in all settings (20 times in 200 articles). The text written in the expected language contains more annotations on average than the texts written by the other models. For every 1000 tokens, there are on average 40 (French, prompt A), 38 (French, prompt B), and 59 (Dutch, prompt A) annotations, compared to 25 average across the other models. There are especially many *strangely/wrongly used word/phrase* annotations, and, in Dutch, a lot of *word/phrase does not exist* annotations (11 such annotations per 1000 tokens). The proportion of those annotations where an influence of English is expected is not much higher than for the other models: 10% clearly suspected influence and 65% no suspected influence on average in French, and 20% and 66% respectively in Dutch with prompt A. Since Zephyr is much worse than the other two models, the following analyses focus mainly on GPT-4 and GEITje.

GPT-4 vs. GEITje (Dutch): Both GPT-4 and GEITje perform a lot better than Zephyr, with fewer annotations on average and fewer texts written in English. The average number of annotations per 1000 tokens (per category) can be seen in Table 4, as well as the proportion of annotations where an influence from English is suspected. When comparing the performance in Dutch of GPT-4 and GEITje (with recommended temperature of .2), a few interesting observations can be made. First, despite GEITje's fine-tuning on Dutch, the experimental setting in Dutch with fewest linguistic annotations was using GPT-4 with Prompt A, though closely followed by GEITje with Prompt A. When non-linguistic remarks are included, GEITje falls further behind. This leads us to a first tentative conclusion regarding this comparison: fine-tuning on Dutch has improved the linguistic quality of GEITje such that it can compete with a much larger LLM like GPT-4 that is not specialised in Dutch. The fact that GEITje is based on the same model (Mistral) as Zephyr, which performs much worse, further strengthens this conclusion. However, the overall non-linguistic quality of texts written by GEITje is not comparable to GPT-4 yet. This is not just reflected in the explicitly annotated *non-linguistic remarks*, but also in the comments shared by the annotators, e.g., about how repetitive the articles written by GEITje are.

English vs. French (GPT-4): Another observation is that there are many more annotations in GPT-4's texts written in French than in Dutch, but, as discussed in the previous section, this can largely be attributed to a disagreement between the Dutch and French annotators on how quickly to use the category *strangely/wrongly used word/phrase*. Considering some room for annotator disagreement in the cross-lingual analysis, it is actually remarkable how similar the average number of annotations are per category in both languages. In terms of the suspected influence of English, more research is needed with cross-lingual comparisons, but this influence appears more present in Dutch than in French. In Dutch, there are proportionally slightly more annotations with a clear suspected influence from English, and one text written completely in English instead of Dutch. This is in line with the findings for Zephyr, though with a better average (linguistic) quality.

Prompt A vs. Prompt B: Across models, the elaborate prompt (A) leads to linguistically better results than the concise prompt (B), but the difference is not always significant. It is most striking for Zephyr in Dutch, where prompt B leads to 36/50 texts written completely in English, and prompt A prevents this from happening in all but 2 cases. The two times where texts were written completely in English by the other two models was also with the concise prompt. For GPT-4 and GEITje respectively, there are on average 1.59 and 3.81 more linguistic annotations per 1000 tokens for the same experiments with prompt B instead of A in Dutch.

The difference is even smaller for French at 0.75. This influence does not appear to affect any specific type of annotations more than the others, and though the general improvement with the elaborate prompt is consistent, it is not statistically significant according to a paired t-test.

Temperature: Because the recommended temperature for GEITje (.2) is much lower than that of Zephyr (.7) and GPT-4 (1.0), GEITje was also tested with a higher temperature. This substantially increased the number of linguistic and non-linguistic annotations. There are significantly (paired t-test, $p < 0.05$) more *strangely/wrongly used word/phrase* annotations with the higher temperature. It is also the setting with most non-linguistic annotations per 1000 tokens out of all, and annotators comment more about strange hallucinations in this setting. The influence of English does not appear affected by the temperature. A notable example of nonsensical output was the following response to prompt A: "Dit is niet mogelijk, aangezien ik een AI-assistent ben die geen Nederlands spreekt." The English translation of this response in Dutch reads: "This is impossible, since I am an AI assistant who does not speak Dutch."

Influence of English: Averaged over all settings, 16% of the annotations are labelled as clearly influenced by English. No influence was suspected for 61% of the annotations. There are big differences per setting and category, but since there are sometimes only a few annotations of a category in a setting, the analysis is limited to those where the differences are large and consistent enough to indicate possible generalisation. Curiously, when averaging over all categories, GEITje displays both the highest and lowest percentage of annotations with a clearly suspected influence from English: 30% for prompt A and the recommended low temperature, versus only 4-5% in the other two settings. On closer inspection, the higher number appears due to a few repeated instances that have a big effect because GEITje's texts tend to be short and contain few annotations. For instance, in one text, *Tour of California* is repeated six times and consistently tagged as clearly influenced by English.

Apart from such cases, the overall influence of English in texts generated by GEITje does appear less obvious than with the other models. Analysing this influence per category results in a few more interesting observations.

The *English word/phrase* annotations regularly concern words that are also often used by native speakers of Dutch and French, except for the texts generated by Zephyr, and the French texts generated by GPT-4, where an average of 75% of those annotations are labelled as not generally used in French or Dutch. This is much lower in the other experiments (combined average of 16%). With grammar and spelling mistakes, there is very little suspected influence of English (on average only 3% with a clear reported influence).

A larger percentage is seen for the *word/phrase does not exist* category (see also the section on IAA for a discussion about this category). Zephyr "makes up" a lot of words, with up to 10.7 such annotations per 1000 words in Dutch using prompt A. Some of the annotations in this category consist of seemingly literal "translations" of English words or phrases. For instance, when referring to traffic congestion, the Dutch word *verkeerscongestatie* is used, which does not exist (0 hits when Googling this word). The first part, *verkeer*, is a correct equivalent of *traffic*. The *s* is correctly added for a correct compound. but *congestatie* is an adaptation of *congestion* that may look Dutch, but does not exist as such (the equivalent of *congestion* can be *congestie* in some cases, but not *congestatie*). And even if *congestatie* were the correct term in Dutch, the compound of with *verkeer* does not exist. Instead, the word *file* is used to refer to traffic congestion. Similarly, in French the phrase *si vous ne pouvez pas les battre, alors rejoignez-les* is used (from *if you cannot beat them, join them*). This French phrase has been used online before (10 Google hits), but is a clear anglicism.

Other observations: Another noteworthy observation made by the annotators was that the writing was inconsistent. This was true for spelling (e.g., *rechts-extremisme* and *rechtsextremisme* in the same article), vocabulary (e.g., switching between *materieel* and *materiaal* in the same article), and punctuation (e.g., French « and English " quotation marks in the same article). Often, multiple options can be considered correct, but it is good practice to remain consistent within a single text. However, since these models are trained on many different types of texts (the exact training data is not disclosed), and don't necessarily contain information about the boundaries between different texts in the training data, it is not surprising that the output contains some inconsistent writing.

As a concluding remark, it is interesting to see annotators comment on the *stylistic features* of generated texts.

> "Certain stylistic features often demonstrate the intervention of artificial intelligence, such as the logical connectors between parts of articles (*en somme, en conclusion, en conséquence, ...*) which are too obvious, unnatural and which would be more nuanced or subtle in a classic article. What also stands out, for being unnatural, is the emphasis often used to describe a situation, a use of dramatic adjectives to describe a sometimes banal situation in an attempt to add effect, I

guess, but it doesn't work at all."

5. Limitations and Conclusions

This exploratory study is a first step towards documenting and better understanding the linguistic qualities of LLMs when writing in Dutch and French, with special focus on the common English bias that is due to the relative overrepresentation of English in the training data o most LLMs. To this end, articles were generated by three different models based on real newspaper headlines, and the resulting corpus was annotated by professional translators for linguistic errors and peculiarities.

Model, language, prompt, and temperature all have a clear impact on results. The difference is noticeable when looking at a simple surface measure like type/token ratio, which is especially high for GEITje, despite repetitive texts. Zephyr is clearly outperformed by the other two models. The most striking result of Zephyr is the number of texts written completely in English instead of Dutch, and the fact that out of the 100 articles to be written by Zephyr based on French prompts French, four were written in English, two in German, and one in Spanish. Linguistically, both GPT-4 and GEITje perform much better and show relatively similar results, indicating that fine-tuning on a specific language can compensate for a smaller model in terms of linguistic quality.

Cross-lingual analyses indicate that the linguistic quality is better in French than Dutch. Comparing a concise and a more elaborate prompt reveals an increased linguistic quality for the latter, though the size of the impact varies per model. Increasing GEITje's very low recommended temperature reduces linguistic quality and increases the number of non-linguistic remarks.

The influence of English is clearly seen for 16% of the annotations on average and can be illustrated very clearly when words or phrases appear to be literally translated from English into Dutch or French words or phrases that are (almost) never used by native speakers.

The main limitations of this study are (1) its scale (limited amount of data per experimental setting), (2) the limited number of languages (only well-resourced languages that are closely related to English), and (3) the potential ambiguity of the annotations. However, the findings can help to narrow down research questions and improve methodologies for experiments on a larger scale. The annotation scheme should be refined to reduce the ambiguity and allow more cross-lingual comparisons.

Since some findings were already relatively clear even with the current setup (e.g., positive impact of elaborate prompt, especially for smaller model), future research can focus more on, e.g., cross-lingual experiments or fine-grained comparison of annotation categories. Given these improvements, expanding the experiments to include more languages will help to improve our understanding of the linguistic qualities of this influential technology. Another worthwhile direction for future research would be to expand the experiments to include more diverse (and perhaps less formal) text types, as the current setup only covered news articles. Further research could also be dedicated to relating and comparing these findings to human linguistic transfer. Knowing whether the influence of human L1 on L2 is similar to the English bias exhibited by LLMs can help to better understand and predict the performance of LLMs.

6. Bibliographical References

Luna De Bruyne. 2023. The Paradox of Multilingual Emotion Detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466. Association for Computational Linguistics.

Lieven De Marez, R Sevenhant, F Denecker, A Georges, G Wuyts, and D Schuurman. 2024. Imec.digimeter.2023. Digitale trends in Vlaanderen.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, prefix=de las useprefix=false family=Casas, given=Diego, 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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There

Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting Linguistic Patterns in Human and LLM-Generated Text.

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, Mo 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, 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 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 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, CJ 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. 2023. Gpt-4 technical report.

Edwin Rijgersberg and Bob Lucassen. 2023. Geitje: een groot open nederlands taalmodel.

Job Schepens, Nicole Marx, and Benjamin Gagl. 2023. Can we utilize Large Language Models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-ofthought reasoners. In *The Eleventh International Conference on Learning Representations*.

Anders Søgaard. 2022. Should We Ban English NLP for a Year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260. Association for Computational Linguistics.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, 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.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Bram Vanroy. 2023. Language Resources for Dutch Large Language Modelling.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. ArXiv:2402.10588 [cs].

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism?

7. Appendix

7.1. Original articles from VRT & RTBF

Tables 5 and 6 list the original articles from the websites of VRT (Flemish) and RTBF (Belgian French) respectively.

nr	title	author	pub. date
1	Vogelgriep treft voor het eerst ijsbeer op Noordpool: "Hier hebben we geen handleiding voor"	Stien Schoofs	03/01/2024
2	Taiwan ontdekt drie Chinese ballonnen in de buurt van luchtmachtbasis	Veerle De Vos	03/01/2024
3	Drie Belgische drugsuithalers opgepakt in Rotterdamse haven, jongste amper 14 jaar	Victor Van Driessche, Belga	03/01/2024
4	Onderzoekers gaan kwab-alen tellen in ondergelopen weides aan Grote Nete	Radio 2, Mathieu Verstichel	03/01/2024
5	Wil je echt vermageren? Zeg dan niet "350 kcal", maar wel "een halfuurtje fietsen"	Dominique Fiers	02/01/2024
6	Rector universiteit Harvard stapt op na ophef over aanpak van antisemitisme en beschuldiging van plagiaat	Nils Schillewaert	02/01/2024
7	Deel van parcours in Gullegem staat onder water: "Maar de veldrit komt niet in het gedrang"	not mentioned	03/01/2024
8	Waarom vond je Belgische tomaten in de winkelrekken op reis in Spanje en Griekenland?	Dennis van den Buijs	03/01/2024
9	Opnieuw miljoenen extra fietsers geteld in provincie Antwerpen: "Alle overheden samen moeten moordstrookjes aanpakken"	Radio 2, Mathieu Verstichel	03/01/2024
10	"Schommelmoment" van verkeersanker Mona krijgt trofee voor mooiste Radio2-moment van 2023	Radio 2, Martijn Donné	02/01/2024
11	Vliegtuigje neergestort tegen geparkeerde auto in Spa: piloot en inzittende overleden	Belga, Kirsten Sokol	28/01/2024
12	New York Times: "Tijdelijk staakt-het-vuren in Gaza van twee maanden in de maak"	Freek Willems	28/01/2024
13	Oekraïense geheime dienst ontdekt fraude bij wapenaankoop, bijna 37 miljoen euro verdwenen	Freek Willems	28/01/2024
14	Tien landen schorten financiering VN-agentschap UNRWA op na beschuldigingen over betrokkenheid bij terreuraanval Hamas	Kirsten Sokol, Joris Truyts, Freek Willems	27/01/2024
15	Van Taylor Swift over Celine Van Ouytsel tot Emma Watson: "deepnudes" overspoelen het internet (en niet alleen op X)	Maarten Bockstaele	28/01/2024
16	Waarom de landbouwers in Europa en bij ons actievoeren	not mentioned	28/01/2024
17	Frans gerecht verklaart acteur Alain Delon beperkt handelingsonbekwaam	Lina El Bakkali, Belga	28/01/2024
18	Intermittent fasting blijft een hype, werkt het ook?	Radio 1, Maxine Rappé	28/01/2024
19	Koning Charles III maakt het goed na zijn prostaatbehandeling	Lukas Lecluyse	26/01/2024
20	Meer vaders nemen een halve dag per week ouderschapsverlof: "Heeft minder impact op je werkweek en op je loon"	Sandra Cardoen	27/09/2023
21	Sport- en energiedrankjes Prime zijn hype bij jongeren, maar hoe ongezond zijn ze?	Wim De Maeseneer, Nils Schillewaert	04/08/2023
22	Klassieke muziek verbindt ons: zelfs onze hartslag synchroniseert	Radio 1, Maxine Rappé	10/11/2023
23	Minister Tinne Van der Straeten ziet geen reden om snel over nieuwe abortuswet te stemmen: "Thema verdient beter"	Joris Truyts, Nils Schillewaert	27/01/2024
24	Duizenden deelsteps verdwijnen uit Brusselse straatbeeld	BRUZZ, Emmanuel Vanbrussel	23/01/2024
25	Nog drie weken tot oudejaar, maar we weten het nu al zeker: 2023 wordt warmste jaar ooit gemeten	Vincent Merckx	06/12/2023
26	Yana's (21) eetstoornis verergerde door TikTok: bijna helft van jongeren ziet berichten over diëten en mager zijn	Dorien Vanmeldert	07/10/2023

nr	title	author	pub. date
27	Apple stoot Samsung na 12 jaar van de troon als grootste smartphoneverkoper ter wereld	Lukas Lecluyse	17/01/2024
28	Oudste bos ooit van 385 miljoen jaar oud strekte zich uit over 400 kilometer	Michaël Torfs	13/01/2024
29	Opnieuw tienduizenden Duitsers op straat tegen uiterst rechts	Joris Truyts, Belga	27/01/2024
30	Batopin vindt moeilijk locaties voor geldautomaten: “Alle suggesties zijn welkom”	Radio 2, Fred Breuls, Bente Vandekeybus	30/01/2024
31	"Hatsjie": het hooikoortsseizoen is begonnen, ontdek op onze pollenbarometer welke pollen je moet vrezen	Vincent Merckx, Belga	30/01/2024
32	Twee slachtoffers door storm Isha in Verenigd Koninkrijk, tienduizenden huishoudens zonder stroom in Ierland	Ellen Maerevoet, Maarten Bockstaele, Sara Van Poucke, Belga	22/01/2024
33	Tot -48 graden (en het voelt nóg kouder): Vlamingen getuigen over ijzige kou in Canada	Zico Saerens	13/01/2024
34	22 Genkse basisscholen hebben eigen bibliotheek: “We willen duidelijk maken dat lezen overal kan”	Radio 2, Fred Breuls	22/12/2023
35	CHECK - Ja, een loonsverhoging levert op voor de staatskas, zoals PS-voorzitter Paul Magnette zegt, maar er zijn ook extra kosten	Nele Baeyens, RTBF, Dorien Vanmeldert	23/01/2024
36	22-jarige Van Uden klopt Groenewegen en Merlier op weg naar eerste sprintzege	not mentioned	30/01/2024
37	Neuralink plaatst eerste hersenimplantaat in menselijk proefpersoon: "We staan nog veraf van hacken van gedachten"	Chris Van den Abeele, Belga, Pieterjan Huyghebaert	30/01/2024
38	Baby "van nog geen uur oud" gevonden in boodschappentas in Londen	Freek Willems	19/01/2024
39	Brand verwoest al bijna 600 hectare van beschermd natuurpark in Argentinië	Lina El Bakkali, Belga	28/01/2024
40	Wilm Vermeir verkozen tot Ruiter van het Jaar, ook zijn paard IQ van het Steentje valt in de prijzen	niet vermeld	16/01/2024
41	Wallonië spendeert per inwoner 70 procent meer aan openbaar vervoer dan Vlaanderen	Rik Arnoudt	27/01/2024
42	Met ChatGPT en geleende spikes: het knotsgekke olympische succesverhaal van John Heymans	Sporza	29/01/2024
43	Mali, Burkina Faso en Niger trekken zich terug uit ECOWAS-verbond	Maarten Bockstaele	29/01/2024
44	Japanse maanlander werkt opnieuw, meer dan een week na de landing	Kathleen Heylen	29/01/2024
45	Eén dode bij aanval van gewapende en gemaskerde mannen in kerk in Istanbul	Joris Truyts	28/01/2024
46	Pakistan voert luchtaanvallen uit op Iran, vrees voor escalatie in de regio	Sara Van Poucke, Nils Schillewaert	18/01/2024
47	Na 2 jaar zicht op nieuwe regering in Noord-Ierland, mét voor het eerst premier van Sinn Féin	Freek Willems	30/01/2024
48	Tomorrowland maakt line-up bekend: op de affiche onder meer David Guetta, Dimitri Vegas & Like Mike en Amber Broos	Belga	25/01/2024
49	Amerikaanse krant The New York Times klaagt OpenAI en Microsoft aan, omdat ze miljoenen artikels gebruikt hebben om ChatGPT te trainen	Wim De Maeseneer, Belga	27/12/2023
50	Drugsdealer loopt tegen de lamp in Brussel, probeert agenten in burger drugs te verkopen	Radio 2, Evi Walschaers	30/01/2024

Table 5: VRT articles

nr	title	author	pub. date
1	Grippe aviaire : un ours polaire infecté en Alaska, une première	Johanne Montay	08/01/2024
2	Taïwan : à quatre jours des présidentielles, le lancement d'un satellite chinois provoque des messages d'alerte	La rédaction, Belga	09/01/2024
3	Rotterdam : arrestation d'un baron de la drogue recherché par la Belgique	Belga, Alain Lechien	05/01/2023
4	Ecraser les oursins violets au marteau pour sauver l'écosystème marin en Californie	Laurick Ayoub sur base d'un reportage de Philippe Jacquemotte	28/12/2023
5	Pourquoi faut-il continuer à faire du sport en hiver ?	Aurélien David via La Une	20/11/2023
6	Suite à plusieurs polémiques, la présidente d'Harvard annonce sa démission	La rédaction	02/01/2024
7	Michael Vanthourenhout s'impose en solitaire à Gullegem en l'absence du "Big Three"	Jâd El Nakadi avec Belga	06/01/2024
8	Selon l'observatoire des prix, 60% des produits alimentaires coûtent moins cher en Belgique qu'ailleurs	QR l'actu	08/01/2024
9	Liège : mauvais bilan 2023 en matière de progrès pour la mobilité cyclable	Marie Bourguignon	02/01/2024
10	Julie Compagnon, les habitants de Bertrix et... la police ont explosé les décibels pour Viva for Life	Par Viva for Life via La Une	22/12/2023
11	Spa : deux morts dans le crash d'un petit avion de tourisme près de l'aérodrome	Olivier Genon	28/01/2024
12	Guerre au Proche-Orient : de violents affrontements sont en cours aux abords des deux principaux hôpitaux de Khan Younès à Gaza	Par La rédaction Info avec Belga	27/01/2024
13	Détournement de 40 millions de dollars par des responsables militaires et chefs d'entreprise ukrainiens	Par La rédaction Info avec Belga	28/01/2024
14	Guerre Israël-Gaza : l'aide à l'Unrwa déjà suspendue par sept pays	Par la rédaction avec AFP	27/01/2024
15	"Protégez Taylor Swift" : les fans se mobilisent pour la défendre contre des deepfakes pornographiques	Par Eléna Lefèbvre	26/01/2024
16	Que compte faire le monde politique en réponse au mécontentement des agriculteurs ?	BELGA – ERIC LALMAND	28/01/2024
17	France : Alain Delon placé sous sauvegarde de justice	Par la rédaction avec AFP	28/01/2024
18	Pour perdre du poids, mieux vaut prendre son petit-déjeuner à 11 heures	Par RTBF avec AFP	20/06/2022
19	Royaume-Uni : le roi Charles III quitte l'hôpital après une opération de la prostate	Par la rédaction avec AFP	28/01/2024
20	Le congé parental n'a jamais été aussi populaire qu'en 2023 en Belgique	Par la rédaction avec Belga	28/01/2024
21	Troubles du sommeil : les boissons énergisantes mises en cause, même à petites doses	Par RTBF avec ETX	28/01/2024
22	La pratique d'un instrument de musique et du chant améliorerait la santé cérébrale des personnes âgées	Par ETX Daily Up édité par Céline Dekock	30/01/2024
23	Avortement : le chantage conservateur du CD&V	Par Philippe Walkowiak	30/01/2024
24	Trottinettes partagées à Bruxelles : Uber et Voi, opérateurs recalés, attaquent la Région en justice	Par Karim Fadoul	30/01/2024
25	Le record de température de 48,8°C en Europe continentale confirmé par l'ONU	Par Marine Lambrecht	30/01/2024
26	Legging legs : la nouvelle tendance controversée et dangereuse qui glorifie la maigreur	Par RTBF avec ETX	30/01/2024
27	Apple dépasse Samsung pour la première fois sur le marché des smartphones	Par Anthony Mirelli	17/01/2024

nr	title	author	pub. date
28	Des scientifiques pensent avoir découvert la plus vieille forêt du monde	Par RTBF Tendance avec AFP	22/12/2019
29	Des milliers de personnes manifestent à nouveau contre l'extrême droite en Allemagne	Par la rédaction avec Belga	27/01/2024
30	La Belgique maintiendra l'accessibilité au cash et aux agences bancaires	Par Maud Wilquin	25/01/2024
31	Les premiers pollens de l'année sont arrivés : la saison des allergies a officiellement commencé	Par Marine Lambrecht	30/01/2024
32	Tempête Isha : un mort en Ecosse, fortes perturbations en Irlande	Par la rédaction avec Belga	22/01/2024
33	Une vague de froid fait au moins 50 morts aux États-Unis	Par la rédaction info avec Belga	20/01/2024
34	20 minutes de lecture obligatoire, tous les vendredis, au lycée François de Sales à Gilly	Par Simon Gerard	30/01/2024
35	Une augmentation des salaires de 2% permet-elle de réduire le déficit de l'État de deux milliards, comme l'affirme Paul Magnette ?	Par Grégoire Ryckmans avec nws check VRT	23/01/2024
36	Casper van Uden surprend Dylan Groenewegen et Tim Merlier sur la première étape de l'AlUla Tour	Par Cédric Lizin	30/01/2024
37	Elon Musk annonce que Neuralink a posé son premier implant cérébral	Par La rédaction avec AFP	30/01/2024
38	Un bébé de moins d'une heure retrouvé vivant dans un sac de courses à Londres	Par rédaction avec AFP	19/01/2024
39	Argentine : un incendie détruit 600 hectares d'un site Unesco	Par Belga	27/01/2024
40	EquiGala : Wilm Vermeir élu cavalier de l'année	Par Louis Lamote	16/01/2024
41	Philippe Henry (Ecolo) : un nouveau contrat de gestion pour les transports en commun, en plein déploiement en Wallonie	Par Par Olivier Arendt, d'après une interview de Thomas Gadisseux via La Première	18/01/2024
42	John Heymans pulvérise le record de Belgique du 5000m indoor et se qualifie pour les Jeux	Par Belga (édité par Alice Devilez)	27/01/2024
43	Les régimes militaires du Burkina, Mali et Niger décident de se retirer de la Cedeao	Par La rédaction Info avec AFP	28/01/2024
44	Le module lunaire japonais a repris vie, les analyses scientifiques vont pouvoir commencer	Par RTBF avec AFP	28/01/2024
45	Une personne décédée lors d'une attaque contre une église catholique italienne à Istanbul	Par La rédaction Info avec AFP	28/01/2024
46	Tensions entre le Pakistan et l'Iran : un problème local aiguisé par le climat régional	Par Pascal Bustamante	18/01/2024
47	Brexit : fin du blocage politique en vue en Irlande du Nord, après deux ans de paralysie	Par la rédaction avec Belga	30/01/2024
48	David Guetta et Swedish House Mafia enflammeront Tomorrowland 2024	Par Belga avec RTBF Culture	26/01/2024
49	Atteinte aux droits d'auteur : le New York Times attaque en justice OpenAI, l'entreprise créatrice de Chat GPT	Par AFP	28/12/2023
50	Plusieurs actions menées par la police à Yser pour limiter le trafic de stupéfiants	Par Belga	30/01/2024

Table 6: RTBF articles

7.2. Other IAA confusion matrices

Tables 7 and 8 represent the inter-annotator agreement matrices between annotators A and C, and B and C respectively. Agreement between A and B was already shown in Table 3. Annotator B was the main annotator.

annotator A → vs C ↓	English word/phrase	grammar mistake	longer piece of English text	non-linguistic remark	other linguistic remark	spelling mistake	strange/wrong construction	strangely/wrongly used word/phrase	word/phrase does not exist	#NA	Total
English word/phrase	14	0	0	0	0	0	0	0	0	0	**14**
grammar mistake	0	21	0	0	0	0	0	0	0	0	**21**
longer piece of English text	0	0	2	0	0	0	0	0	0	0	**2**
non-linguistic remark	0	0	0	2	0	0	0	0	0	2	**4**
other linguistic remark	0	2	0	2	5	0	0	0	0	1	**10**
spelling mistake	0	0	0	0	0	11	0	0	0	5	**16**
strange/wrong construction	0	1	0	0	0	0	25	0	0	5	**31**
strangely/wrongly used word/phrase	0	0	0	0	0	0	1	29	0	4	**34**
word/phrase does not exist	0	0	0	0	0	0	0	0	16	0	**16**
#NA	1	2	0	1	0	8	2	2	0	23	**39**
Total	**15**	**26**	**2**	**5**	**5**	**19**	**28**	**31**	**16**	**40**	**187**

Table 7: Confusion matrix between annotators A and C

annotator B → vs C ↓	English word/phrase	grammar mistake	longer piece of English text	non-linguistic remark	other linguistic remark	spelling mistake	strange/wrong construction	strangely/wrongly used word/phrase	word/phrase does not exist	#NA	Total
English word/phrase	13	0	0	0	0	1	0	0	0	0	**14**
grammar mistake	0	19	0	0	0	0	0	0	0	2	**21**
longer piece of English text	0	0	2	0	0	0	0	0	0	0	**2**
non-linguistic remark	0	0	0	1	3	0	0	0	0	0	**4**
other linguistic remark	0	2	0	0	8	0	0	0	0	0	**10**
spelling mistake	0	1	0	0	0	12	0	0	1	2	**16**
strange/wrong construction	0	3	0	0	0	0	24	2	0	2	**31**
strangely/wrongly used word/phrase	0	2	0	0	0	0	1	27	1	3	**34**
word/phrase does not exist	2	0	0	0	0	0	0	0	14	0	**16**
#NA		10	0	0	2	5	9	4	1	8	**39**
Total	**15**	**37**	**2**	**1**	**13**	**18**	**34**	**33**	**17**	**17**	**187**

Table 8: Confusion matrix between annotators B and C