Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

The Effect of Word Predictability on Spoken Cross-Language Intelligibility

Wei Xue[1,2], *Iuliia Zaitova*[2], *Bernd Möbius*[2]

[1]Frontier Institute of Science and Technology, Xi'an Jiaotong University, China
[2]Department of Language Science and Technology, Saarland University, Germany

weixue@lst.uni-saarland.de, izaitova@lsv.uni-saarland.de, moebius@lst.uni-saarland.de

Abstract

Cross-language intelligibility refers to how well speakers of language A understand language B without prior learning. While the impact of linguistic and extra-linguistic factors on cross-language intelligibility has been widely studied, the effect of word predictability, known to impact comprehension and speech perception, remains underexplored. This study examines this effect by comparing German and English native speakers translating Dutch words presented in Dutch spoken sentential utterances with varying word predictability. We also investigate whether additional written context would aid cross-language intelligibility. Our results showed that word predictability significantly influences cross-language intelligibility, with German speakers experiencing even stronger effects, whereas only English speakers benefit from the additional written context. These findings suggest that word predictability dynamically shapes cross-language intelligibility, tending to be language-specific.

Index Terms: cross-language intelligibility, predictability, speech perception, receptive multilingualism

1. Introduction

Receptive multilingualism (RM) is a mode of communication where speakers of language A understand utterances in language B without prior learning [1]. The success of RM can be evaluated by examining mutual intelligibility (the average of bidirectional comprehension between languages A and B) and cross-language intelligibility (unidirectional comprehension from language A to language B), and is expected to be low for distant languages like English and Chinese. Previous research has extensively explored linguistic factors (e.g., lexical and phonetic distances) and extra-linguistic factors (e.g., attitude, exposure) that contribute to intelligibility [2, 3, 4, 5].

While linguistic and extra-linguistic factors affecting cross-language intelligibility have been widely studied, the role of word predictability for cross-language intelligibility remains underexplored. Word predictability, i.e., the likelihood of a word being anticipated based on its syntactic, semantic, and pragmatic context, has been shown to enhance speech perception and language comprehension in native language processing [6, 7]. For instance, the word *houses* is highly predictable in the sentence *People live in houses* but much less so in *He turned and saw the houses*. If similar mechanisms extend to RM, greater predictability may facilitate cross-language intelligibility.

Previous studies have indirectly touched on this issue. For example, Gooskens [1] used cloze tests, where participants filled in blanks in a paragraph by selecting from a list of word candidates, demonstrating that surrounding context likely aided comprehension. However, the precise influence of word predictability has not been systematically investigated. While some other studies used large language models to quantify word unexpectedness [4], our study employs an experimentally controlled validation of stimuli that combines automatic measures with human judgments of word predictability to evaluate the direct impact of word predictability on intelligibility.

Additionally, contextual cues (leading to higher/lower word predictability) were suggested by several models of speech perception to interact with phonetic similarity during word recognition. The TRACE model [8] posits interactive processing across features, phonemes, and words, while the Fuzzy Logical Model of Speech Perception (FLMP) [9] suggests that recognition is optimized by integrating multiple probabilistic sources of information, including auditory, visual, and contextual cues.

Despite the focus of these models on native and second-language processing, their implications for cross-language word recognition in RM remain unclear. In cross-language intelligibility, the complexity of word recognition increases, as speakers of language B must process acoustic input from language A and map it onto their own linguistic representations. When the word-level representations between two languages are highly similar, additional cues (such as context and modality) may enhance recognition, as speakers of language B recognize the contextual cues presented in language A, akin to what is observed in native language processing. However, as linguistic distance increases, word predictability facilitation might be reduced.

In this study, we set up a free translation experiment comparing German and English native speakers translating Dutch words presented in Dutch spoken sentential utterances with varied degrees of word predictability (i.e., a word being more or less predictable given preceding context). We are interested in filling the gap of how word predictability affects cross-language intelligibility. We also investigate whether additional written context would facilitate word-level cross-language intelligibility. Further, by comparing the German and English speakers' translation performance, we aim to explore whether the effect of word predictability is language-specific.

2. Method

2.1. Stimuli

We selected 15 target words that are cognates in the three Germanic languages (i.e., Dutch, German, and English). These cognates have word lemma frequencies higher than 20/million in CELEX [10] and 10/per million in SUBTLEX [11, 12, 13].

To control the predictability of our cognates in their immediate context, for each cognate in Dutch, we extracted one word-based trigram with a high surprisal value (i.e., high unexpectedness given preceding context), which is always a preposition phrase, and one with a low surprisal value, which is always a

10.21437/Interspeech.2025-1676

Table 1: *Example of a paired sentence given a selected word-based trigram of the target word. English translations of the sentences are in brackets.*

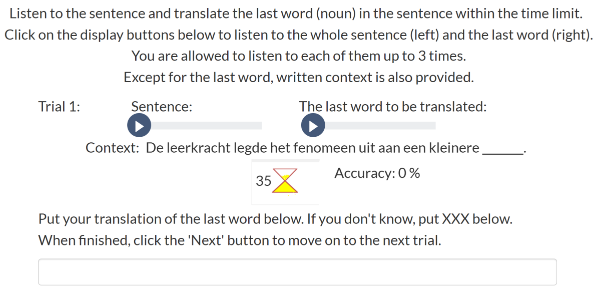
Word predictability	Trigram surprisal	Sentence example
Pred	High	De jongen raakte de bal met de **arm**. ("The boy touched the ball with the **arm**.")
Unpred	High	Hij maakte een mooie beweging met de **arm**. ("He made a nice movement with the **arm**.")
Pred	Low	Hij masseerde zachtjes zijn andere **arm**. ("He gently massaged his other **arm**.")
Unpred	Low	Ze toonde trots zijn andere **arm**. ("She proudly showed his other **arm**.")

noun phrase. We used trigram because it offers a better balance between capturing meaningful linguistic patterns and avoiding overfitting. We further controlled that these trigrams translated to German and English are consistent with high or low surprisal values. These trigrams and their surprisal values were extracted from three monolingual trigram language models, one for each language, trained on CGN [14] for Dutch, ukWaC for English, and deWaC for German [15] following the practice in [16]. Note that we used a relative comparison rather than any absolute thresholds for high or low surprisal values. This is because such thresholds may be different depending on the models used and their training data, and thus difficult to generalize to different stimuli. Additionally, the phrase type (preposition vs. noun phrase) and surprisal levels were tangled due to practical constraints aimed at maintaining cross-linguistic consistency.

Lastly, we embedded each trigram into two manually constructed Dutch sentences, one in which the target word was highly predictable from the preceding context and the other where it had a low predictability, resulting in four sentences per cognate. We ensured that when translating these sentences into German and English, the cognates always appeared in sentence-final position to minimize syntactic and grammatical confounds. The translated sentences were verified by one German and one English native speaker, who are also professional linguists. In total, we generated 60 experimental sentences, categorized into four subsets based on word predictability (given preceding sentential context) and trigram surprisal. Table 1 provides an example using the cognate "arm". The mean (SD) for the number of words in the context are 7.3 (1.3) for those with low-surprisal trigrams and 6.7 (0.8) for high-surprisal trigrams.

Before our free translation experiment, we validated the stimuli by combining automatic measures with human judgments. We paired 60 sentences by trigram for each cognate and asked 32 Dutch native speakers (a payment of €12/h, age under 55, gender-balanced) from Prolific (https://www.prolific.com/) to choose the sentence that best fit the last word (the cognate). Surprisal values for cognates were obtained from the pre-trained Dutch language model GroNLP/gpt2-small-dutch[1] [17]. A moderate, significant correlation was found between the difference in response preference and that in surprisal between paired sentences (Spearman's $r = 0.47$, $p < .05$, an alpha level used for statistical significance). Three pairs for low-surprisal and four for high-surprisal trigrams showed no clear preference (difference < 20) and were used as fillers, excluding them from statistical analyses in Section 2.4.

After validating the stimuli, we made recordings of sentences from a female Dutch native speaker (age=27) in a self-paced reading session with randomized sentences. The recordings were made in a sound-attenuated booth at 44.1 kHz and later resampled to 16 kHz with intensity adjusted to 70 dB. We extracted cognates from the corresponding sentence recordings for use in subsequent experiments described in Section 2.2.



Listen to the sentence and translate the last word (noun) in the sentence within the time limit.
Click on the display buttons below to listen to the whole sentence (left) and the last word (right).
You are allowed to listen to each of them up to 3 times.
Except for the last word, written context is also provided.

Trial 1: Sentence: The last word to be translated:

Context: De leerkracht legde het fenomeen uit aan een kleinere ______.

35 Accuracy: 0 %

Put your translation of the last word below. If you don't know, put XXX below.
When finished, click the 'Next' button to move on to the next trial.

Figure 1: *Screenshot of the free translation task in the AudioText type of experiments for English participants.*

Table 2: *Participant counts (female/male) in experimental conditions across languages. Numbers in parentheses show female/male distribution. Total participant counts include non-binary individuals not shown in ratios.*

Language	Word	Trigram	Participants	
	Pred.	Surprisal	AudioOnly	AudioText
German	Pred	Low	17 (8/9)	14 (5/9)
	Unpred	Low	20 (11/9)	12 (2/10)
	Pred	High	19 (9/10)	18 (7/10)
	Unpred	High	19 (9/10)	17 (7/10)
English	Pred	Low	20 (10/10)	21 (11/10)
	Unpred	Low	19 (10/9)	20 (10/10)
	Pred	High	20 (10/10)	20 (10/10)
	Unpred	High	19 (12/7)	20 (10/10)

2.2. Experimental design

We conducted two types of free translation experiment (AudioOnly vs. AudioText) where native German or English speakers were asked to translate the cognate at the end of the utterance within a time limit. The two experiments differ in whether the written text of the sentential context (shown in the sentence examples in Table 1) is presented in addition to the audio clips. A screenshot of the task in the AudioText type of experiment is shown in Figure 1. The allotted time for translation (represented by the yellow hourglass in Figure 1) was the sum of 10 seconds for cognate and 3 seconds per word in its context. Participants were presented with two audio clips: the whole sentence and the cognate. They were allowed to listen to each audio clip up to three times, and they had to listen to the whole-sentence clip, which includes the cognate, at least once to ensure their exposure to the sentential context. Also, to reduce any repetition or memory effect, the 60 sentences were distributed into four subsets in terms of the trigram surprisal (Low vs. High) and the word predictability (Pred vs. Unpred).

2.3. Participants

We recruited around 20 participants per subset per experiment type via Prolific (https://www.prolific.com/) with a compensation of 12€/h. The exact numbers of participants are shown

[1] https://huggingface.co/GroNLP/gpt2-small-dutch/

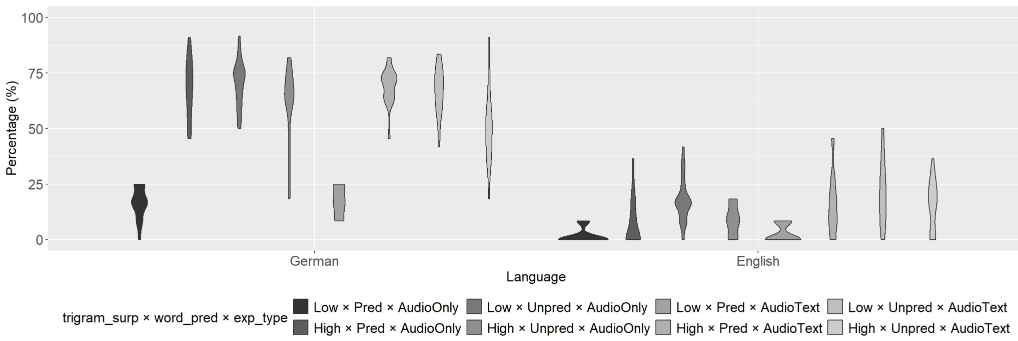


Figure 2: *Percentage of correct translation averaged for participants (i.e., one score per participant) for different levels of trigram surprisal (i.e., trigram_surp with 'Low' and 'High') and word predictability (i.e., word_pred with 'Pred' and 'Unpred'), as well as two types of experiments (i.e., exp_type with 'AudioOnly' and 'AudioText'). For each language, the violin diagrams are organized from left to right in the order of top to bottom and left to right in the legend.*

in Table 2, and they were rather gender-balanced. The participants were adults aged 62 or younger with no reported hearing loss. None had prior experience learning Dutch and had minimal exposure to the language. The ethics committee of Saarland University approved the studies, and all participants provided informed consent before taking part.

2.4. Statistical analyses

We conducted a Generalized Linear Mixed-Effect Model (GLMM) to predict participants' translation accuracy by using the factorial variables. The formula is as follows:

is_correct ∼ exp_type
+ trigram_surp * word_pred
+ (1 | Participant)
+ (1 + trigram_surp + word_pred | cognate)

where the dependent variable is_correct refers to a translation from a participant to be correct (1) or not (0); exp_type refers to the two types of experiments (AudioOnly vs. AudioText); trigram_surp refers to the two trigrams varied in their surprisal (Low vs. High); word_pred refers to the ending cognate word being predictable (Pred) or not (Unpred) in the sentence given its preceding context. We also included a random intercept for participants (Participant) and cognate items (cognate) as well as a random slope for cognate over word_pred and trigram_surp. These two-level factorial variables were dummy coded with 0, AudioOnly, Low, Pred as the reference level for is_correct, exp_type, trigram_surp, and word_pred, respectively. All GLMMs were run with a logit link and were controlled with the bobyqa optimizer and a maximum number of iterations of $2 * 10^5$ to improve model convergence. All statistical analyses were conducted in R [18] by using the *glmer* function in the *lme4* package [19], as well as *ggplot2* [20] for visualization.

Table 3: *Random effects of glmer models for German and English data.*

Language	Groups	Name	Variance	Std.Dev.	Corr	
German	Participant	(Intercept)	0.7118	0.8437		
	cognate	(Intercept)	9.8862	3.1442		
		High	7.0608	2.6572	-0.72	
		Unpred	7.1578	2.6754	-0.69	0.89
English	Participant	(Intercept)	0.4625	0.6801		
	cognate	(Intercept)	1.2521	1.1190		
		High	2.9146	1.7072	-0.93	
		Unpred	3.7678	1.9411	-0.49	0.64

3. Results and discussion

3.1. Descriptive results

We first visualized the participants' responses regarding their accuracy as shown in Figure 2. The accuracy was averaged for participants. Our German participants showed significantly higher accuracy than our English participants[2]. Besides, there is a clear low accuracy for low trigram surprisal (trigram_surp) combined with high word predictability (word_pred), which contrasts with our expectation when constructing the stimuli. This applies to both German and English participants, although the difference in English is smaller than in German.

3.2. GLMMs

As there is a clear distinction between German and English response distributions, we applied GLMMs for German and English data separately. The random effects analysis (Table 3) shows greater variability in accuracy among German participants compared to English participants, both at the individual (0.7118 vs. 0.4625) and item (cognate) levels (9.8862 vs. 1.2521). The larger variances of German participants for the random slope of cognate indicate that they exhibit larger fluctuations in how trigram surprisal and (un)predictability affect word recognition. The negative correlations between the intercept and surprisal/(un)predictability effects suggest that words that are generally easier tend to be less affected by these factors, with this trend being stronger in English (-0.93) than in German (-0.72/-0.69). For instance, in English, the easiest words are even less influenced by trigram surprisal than in German (-0.72 vs. -0.93). Additionally, High surprisal and word (Un)predictability effects are more strongly related in German (0.89) than in English (0.64), indicating that German speakers rely more on contextual cues, making their performance more sensitive to word-specific difficulty.

3.2.1. German data

The fixed effect results for the German data are shown in Table 4. The intercept estimate (-3.1510) and its *p*-value ($< .05$) suggest that the accuracy for the baseline, namely Low trigram surprisal, high word predictability (Pred), and AudioOnly type, is significantly below chance $(P(correct) = \frac{e^{3.1510}}{1+e^{3.1510}} =$

[2] based on the Welch Two Sample t-test of correctness by language (t = 28.671, df = 2548.3, p-value $< .05$)

Table 4: *Fixed effects of glmer models for German and English data. Significance (Sig.) codes: *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.*

Language	Fixed Effect	Estimate	Std. Error	z value	Pr($> \|z\|$)	Sig.
German	(Intercept)	-3.1510	0.9884	-3.188	0.00143	**
	High	4.2579	0.9364	4.547	5.44e-06	***
	Unpred	4.7789	0.8752	5.460	4.75e-08	***
	AudioText	-0.3259	0.2937	-1.109	0.26727	
	High:Unpred	-5.2053	0.6045	-8.611	<2e-16	***
English	(Intercept)	-4.9641	0.5988	-8.290	<2e-16	***
	High	2.3037	0.7115	3.238	0.00121	**
	Unpred	2.5143	0.7344	3.424	0.00062	***
	AudioText	0.6335	0.2155	2.939	0.00329	**
	High:Unpred	-3.4539	0.6723	-5.137	2.79e-07	***

0.0409 (or 4.1%)). It is clear that having high trigram surprisal (High) and low word predictability (Unpred) resulted in significant positive estimates (4.2579 and 4.7789), namely higher log-odds of having correct translations as the models predict 1 (the correct answer). These results contrast with our expectation when designing the stimuli. Recall that our stimuli were validated by Dutch native speakers. The results indicate a substantial difference between native and cross-language comprehension. The AudioText experimental type resulted in lower (-0.3259), but insignificant ($p > .05$) accuracy. We also found a significant interaction between trigram surprisal and word predictability. In detail, when trigram surprisal is High, having low word predictability (Unpred) significantly led to lower accuracy (-5.2053). Overall, high trigram surprisal and a word being unpredictable separately increase accuracy, but their interaction strongly decreases accuracy, suggesting that when both factors are present, they interfere with comprehension.

3.2.2. English data

The fixed effect results for English are shown in Table 4. The baseline accuracy is much lower (0.69%) compared to that (4.09%) in German. Having High trigram surprisal and low word predictability (Unpred) resulted in significantly higher accuracy (2.3037 and 2.5143), but weaker effects of English compared to German. In line with the German results, these results again contrast with our expectation in stimulus design. Further, contrary to the German results, the AudioText type resulted in significantly higher accuracy (*Estimate* = 0.6335, $p < .05$). Also, we observed again a significant interaction between trigram surprisal and word predictability but a less severe effect compared to our German participants (-3.4539 vs. -5.2053).

In summary, our results suggest that even when Dutch words are "predictable" (only based on the validation of the stimuli) and presented in a simple auditory format, participants found them very difficult to understand. The low baseline seems to show that without additional cues, cross-language intelligibility is minimal. The even lower baseline accuracy for our English participants suggests they struggle more with cross-language comprehension than our German participants do. German and Dutch have closer typologically proximity [21], making it slightly easier for German participants. They seem to rely more on predictability effects (trigram surprisal and word predictability), but suffer a larger accuracy drop when both factors (High surprisal and Unpredictability) are present. Whereas English participants benefit more from AudioText but show smaller predictability effects overall, implying that English participants encountered more difficulties in comprehending foreign, spoken utterances and thus benefit more from direct, written forms of context. The interaction of High and Unpred had negative estimates, meaning that when both factors are extreme like the native comprehension in stimulus validation, they impair comprehension. However, as the trigram surprisal levels were tangled with the trigram syntax, future research may disentangle the effects of trigram surprisal and trigram syntax.

3.3. Limitations

Our GLMM findings did not entirely align with predictions from stimulus validation. However, within high-surprisal trigrams, lower word predictability (Unpred) corresponded to decreased translation accuracy. Importantly, the surprisal manipulation was intertwined with syntactic structure, as high-surprisal trigrams consistently appeared as prepositional phrases, while low-surprisal ones were noun phrases, potentially affecting outcomes. Future work could disentangle surprisal from syntactic structures and consider additional factors, such as the effect of participants' L2 proficiency, phonetic distances among items, cultural context, as well as explore using stimuli validated across all three languages. Participants' age may also affect performance, as younger participants could be more adaptable due to greater neuroplasticity, while older ones might depend more on their native language's acoustic and phonetic cues. Hearing ability and cognitive speed may also contribute.

4. Conclusion

In this study, we explored how word predictability affects cross-language intelligibility in comprehending spoken, foreign language and whether additional written text facilitates comprehension. Regardless of participants' native language (English or German), word predictability and trigram surprisal significantly influence cross-language intelligibility (of Dutch stimuli), with lower predictability in combination with higher trigram surprisal decreasing accuracy. This interaction suggests that word predictability benefits diminish under high trigram surprisal conditions. We observed substantial individual variability among participants and across cognate items, with stronger effects observed in German speakers, and that the influence of trigram surprisal and word predictability varies considerably across items, highlighting the complexity of cross-language word recognition. Together, these findings suggest that word predictability dynamically shapes intelligibility in receptive multilingualism, tending to be language-specific.

5. Acknowledgements

This research is supported by a Rubicon grant from the Dutch Research Council (NWO) to the first author (019.223SG.004) and a grant by German Research Foundation (DFG, Project ID 232722074-SFB1102) to the second and third authors. We would also like to thank our colleague Marian Marchal for helping us preparing the stimuli.

6. References

[1] C. Gooskens, *Mutual intelligibility between closely related languages*. Walter de Gruyter GmbH & Co KG, 2024, vol. 30.

[2] K. Jágrová, T. Avgustinova, I. Stenger, and A. Fischer, "Language models, surprisal and fantasy in slavic intercomprehension," *Computer Speech Language*, vol. 53, 06 2018.

[3] I. Stenger and T. Avgustinova, "On Slavic cognate recognition in context," in *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue'*, vol. 20, Moscow, Russia, June 2021, pp. 660–668.

[4] K. Jágrová, M. Hedderich, M. Mosbach, T. Avgustinova, and D. Klakow, "On the correlation of context-aware language models with the intelligibility of polish target words to czech readers," *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.662277

[5] I. Zaitova, I. Stenger, W. Xue, T. Avgustinova, B. Möbius, and D. Klakow, "Cross-linguistic intelligibility of non-compositional expressions in spoken context," in *Proceedings of Interspeech 2024*, Saarbrücken, Germany, September 2024. [Online]. Available: https://www.researchgate.net/publication/383651915_Cross-Linguistic_Intelligibility_of_Non-Compositional_Expressions_in_Spoken_Context

[6] J. Manker, "Contextual predictability and phonetic attention," *Journal of Phonetics*, vol. 75, pp. 94–112, 2019.

[7] M. J. Pickering and S. Garrod, "Do people use language production to make predictions during comprehension?" *Trends in cognitive sciences*, vol. 11, no. 3, pp. 105–110, 2007.

[8] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[9] D. W. Massaro, *Perceiving talking faces: From speech perception to a behavioral principle*. Mit Press, 1998.

[10] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "Celex2 ldc96l14 [web download]," 1995.

[11] E. Keuleers, M. Brysbaert, and B. New, "Subtlex-nl: a new measure for dutch word frequency based on film subtitles," *Behavior Research Methods*, no. 3, 2010.

[12] M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl, "The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in german," *Experimental Psychology*, no. 5, 2011.

[13] D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The english lexicon project," *Behavior Research Methods*, vol. 39, 2007.

[14] I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden, "CGN, an annotated corpus of spoken Dutch," in *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*, 2003. [Online]. Available: https://aclanthology.org/W03-2414

[15] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, "The wacky wide web: a collection of very large linguistically processed web-crawled corpora," *Language resources and evaluation*, vol. 43.3, pp. 209–226, 2009.

[16] O. Ibrahim, I. Yuen, M. van Os, B. Andreeva, and B. Möbius, "The combined effects of contextual predictability and noise on the acoustic realisation of german syllables," *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 911–920, 2022.

[17] W. de Vries and M. Nissim, "As good as new. how to successfully recycle english gpt-2 to make models for other languages," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021, p. 836–846. [Online]. Available: http://dx.doi.org/10.18653/v1/2021.findings-acl.74

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[19] D. Bates, "lme4: Linear mixed-effects models using eigen and s4," *R package version*, vol. 1, p. 1, 2016.

[20] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[21] G. Booij, *The Phonology of Dutch*, ser. The Phonology of the World's Languages. Oxford University Press, 1995.