DOI: 10.30420/456617005

The Role of Surprisal and Entropy in Spoken Intercomprehension: An Experiment on Translation of Cognates with Varied Predictability

Wei Xue†, Julius Steuer†, Dietrich Klakow†, Bernd Möbius†*

*Interdisciplinary Research Center of Frontier Science and Technology, Xi'an Jiaotong University, China
†Department of Language Science and Technology, Saarland University, Germany
Email: wei.xue@xjtu.edu.cn, jsteuer@lsv.uni-saarland.de, dietrich.klakow@lsv.uni-saarland.de, moebius@lst.uni-saarland.de

Abstract

Word predictability affects comprehension and speech perception, yet its role in intercomprehension – understanding foreign languages without prior learning – remains understudied. While surprisal and entropy, derived from language models (LMs), capture different aspects of predictability, this study aims to explore how these estimates explain intercomprehension success. We asked German and English native speakers to translate Dutch cognates from spoken utterances with varied predictability. We extracted the estimates from cascade Automatic Speech Recognition with LM and LM-only approaches. Results revealed that both approaches explained translation accuracy to a similar extent, but only LM-only estimates were significant. Also, German speakers seem to leverage contextual information as in native comprehension, while English speakers do not. These findings highlight that beyond LM-based estimates, typological proximity shapes intercomprehension in varied predictability contexts.

1 Introduction

Intercomprehension refers to the ability of speakers to understand foreign but related languages without prior study by leveraging their existing linguistic knowledge, either in their first (L1) or second (L2) language [1, 2]. This phenomenon is facilitated by linguistic similarities between languages, such as phonemes, lexicon, vocabulary, and grammar. However, beyond these structural similarities, other factors that influence cognitive load and consequently language comprehension may also play a role. One such factor is word predictability [3].

Predictability, the expectation of upcoming words, is a key component of language comprehension [3–5]. It reflects how likely a word is to occur based on contextual information and prior linguistic knowledge. Predictability is high when context strongly constrains the continuation of a sentence (e.g., *houses* in *People live in houses.*), and conversely, low when context provides little guidance (e.g., *houses* in *He turned and saw the houses.*). Extensive research has shown that word predictability contributes to reading [6, 7] and listening comprehension in native and non-native speech [8]. However, its role in intercomprehension, where listeners attempt to understand an unfamiliar but related language, remains largely unexplored.

A useful framework for quantifying predictability comes from statistical language models (LMs), which estimate the likelihood of words in a sequence based on learned linguistic patterns. Two key LM-based estimates – surprisal and entropy – capture different aspects of predictability [9]. Surprisal is the negative logarithm of a word's probability. It measures how unexpected a word is given its preceding context, with higher surprisal indicating lower predictability. Entropy, on the other hand, is the weighted sum over the surprisal of all the words that could appear in the word's position. It quantifies the uncertainty in predicting the next word, with higher entropy suggesting a broader range of plausible continuations. These estimates have been shown to correlate with human comprehension performance, like with self-paced reading times [9], and have been linked to success in intercomprehension tasks, such as translating multiword expressions [10, 11].

Despite these connections, it is unclear how surprisal and entropy explain the ease or difficulty of understanding foreign but related languages, as in intercomprehension, when contextual predictability varies. In this study, we aim to investigate how word predictability influences intercomprehension success by manipulating the contextual predictability of target words in sentences and examining how these two LM-based estimates relate to it.

To this end, we focus on three Germanic languages: Dutch, German, and English. We conduct a free translation experiment where native speakers of German and English listen to Dutch spoken utterances and attempt to translate the final word (cognate) of each sentence. To systematically manipulate word predictability, we modify the preceding sentence context, making the final word either highly or less predictable. We then use LM-based surprisal and entropy as predictors to examine whether they explain variations in translation success, shedding light on the cognitive mechanisms underlying intercomprehension and the role of probabilistic linguistic expectations in this process.

2 Method

2.1 Stimuli

We selected 15 cognates shared across Dutch, German, and English[1], ensuring that their lemma frequencies exceeded 20 per million in CELEX [12] and 10 per million in SUBTLEX [13–15]. For each cognate, we identified one word-based Dutch trigram with high surprisal (a prepositional phrase) and one with low surprisal (a noun phrase). These trigrams were translated into German and English while maintaining their surprisal levels. To ensure linguistic consistency, we extracted the trigrams using monolingual language models trained on CGN [16] for Dutch, ukWaC for English, and deWaC for German [17] following the practice in [18]. Rather than applying absolute surprisal thresholds, we used relative comparisons across models and datasets. Note that the phrase structures were tangled with high/low surprisal values for consistency across languages. Lastly, for each trigram, we created two Dutch sentences where the target word (i.e., the cognate) was either highly or less predictable, resulting in

[1] Note that English has larger phonetic distances to Dutch than German due to the nature of the languages.

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**.)

Table 1: Example of sentences for the target word "arm" (in bold) with selected word-based trigrams (underlined) in four conditions. English translations of the sentences are in brackets.

four sentences per cognate. Also, in the German and English translations of the Dutch sentences, cognates were always placed at the sentence-final position to control for syntactic and grammatical effects. The translations were reviewed by native German and English speakers. Altogether, we generated 60 sentences, categorized into four conditions based on word predictability and trigram surprisal (see Table 1 for an example with "arm").

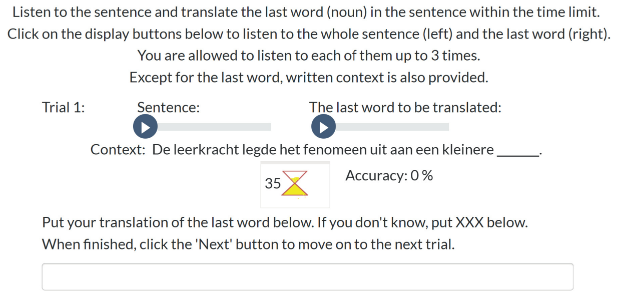
Before conducting the free translation experiment, we validated the stimuli by presenting 30 pairs of sentences, each pair containing two sentence variants of a cognate's trigram, to 32 Dutch native speakers. They indicated which sentence provided a better contextual fit for the final word. These participants were balanced in gender and under 54 years old. The selection preferences showed a moderate, significant correlation with the difference in cognate surprisal, obtained from the pre-trained Dutch language model GroNLP/gpt2-small-dutch[2] [19], between sentence pairs (Spearman $r = 0.47$, $p = 0.022$). After validation, a 27-year-old female Dutch native speaker recorded the sentences in a self-paced reading session. The recordings were made at 44.1 kHz, resampled to 16 kHz and adjusted to 70 dB in Praat [20]. Cognate audios were extracted from the corresponding sentential utterances for use in the experiments detailed in Section 2.2.

2.2 Experimental design

We conducted two versions of free translation experiments (Audio vs. AudioText) via Labvanced [21] where native German or English speakers were asked to translate the cognate at the end of the utterance within a time limit. In both versions of the experiments, participants were presented with two audio clips, one for the whole sentence including the cognate, and the other one with only the cognate extracted from the whole-sentence utterance. They were allowed to listen to each audio clip three times maximum and minimum once to the whole sentence to ensure their exposure to the sentential context. The allotted time for translating one cognate was 10 seconds per cognate and 3 seconds per contextual word represented by the yellow hourglass in Figure 1. The two experiments differ in whether the written text of the sentential context is given in addition to the audio clips, as shown in Figure 1.

2.3 Participants

We recruited our participants via Prolific[3] with a compensation of 12€/h. The number of participants is shown in Table 2. All participants were native speakers of German (Europe) or English (US), aged 18–62, never learned Dutch before, with minimal prior exposure to Dutch and no self-reported hearing loss. The studies were approved by the ethics committee of Saarland University, and all our participants gave their consent before participation.



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 AudioText version of the experiment for English participants.

Language	Word predictability	Trigram surprisal	Nr of participants (Audio / AudioText)
German	Pred	Low	17 / 14
	Unpred	Low	20 / 12
	Pred	High	19 / 18
	Unpred	High	19 / 17
English	Pred	Low	20 / 21
	Unpred	Low	19 / 20
	Pred	High	20 / 20
	Unpred	High	19 / 20

Table 2: Number of participants for German and English in each of the four conditions.

2.4 Surprisal and entropy extraction

To address our research aim, we extracted surprisal and entropy estimates through two approaches: (1) cascaded Automatic Speech Recognition (ASR) system with LM and (2) LM only. The estimates were extracted for (1) cognates given their context and (2) averaged over the whole sentence. The latter were calculated by first summing the values of each word in the sentence given their preceding context, and then dividing by the total number of words. In total, we had eight variables, and they were all scaled to their means. Details are explained below:

(1) cascaded ASR with LM: We first applied monolingual ASR models of German and English to Dutch audios to generate transcriptions and used the generated transcriptions as inputs for monolingual LM models of German and English to extract surprisal and entropy estimates. This approach was to emulate native German and English speakers listening to Dutch spoken utterances.

(2) LM only: In this approach, the inputs to monolingual LMs were sentential contexts in Dutch and cognates in German/English. This approach was to emulate German and English speakers reading Dutch context while being able to recognise the cognates, as participants can experience in the AudioText version of the experiment.

The ASR models were jonatasgrosman/wav2vec2-large-xlsr-53-german[4] for German and jonatasgrosman/

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

[3] https://www.prolific.com/

[4] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german/

wav2vec2-large-xlsr-53-english[5] for English [22]. Both ASR models were based on facebook/wav2vec2-large-xlsr-53 [23] and fine-tuned in German/English using the train and validation splits of Common Voice 6.1. The LM models are Transformer-based, GPT models, meaning that the models see the preceding context all at once in contrast to the traditional n-gram models that see only the nearest n-1 words. The LM models are benjamin/gerpt2[6] for German [24] and openai-community/gpt2[7] for English [25] . We used minicons [26] to extract surprisal values.

2.5 Statistical analyses

To address whether the LM-based estimates can predict cross-language intelligibility (reflected by the correctness of our participants' translations) obtained at different levels of word predictability and trigram surprisal, we conducted generalized linear mixed-effect models (GLMMs) to examine the contribution of our LM-based estimates in addition to other experimental factors. We used *ggplot2* [27] for visualization and *glmer* function in the *lme4* package [28] for GLMMs, both implemented in R [29]. The experimental factors include experiment version (binomial variable exp_version being Audio or AudioText), trigram surprisal (binomial variable trigram_surp being Low or High), and word predictability (binomial variable word_pred being Pred or Unpred) in explaining cross-language intelligibility, reflected by translations being correct or not (binomial variable is_correct). The GLMMs were fitted separately for German and English with the same formula:

is_correct ∼ exp_version + trigram_surp * word_pred
+ **estimateX**
+ (1 | participant_id) + (1 | cognate)

where estimateX is a selection of the eight estimates explained in Section 2.4. The reference levels of the variables are 0 (incorrect) for is_correct, Audio for exp_version, Low for trigram_surp, and Pred for word_pred using dummy coding. We included random intercepts for participants (participant_id) and items (cognate) for both languages but random slopes of High and Unpred for cognate for German due to better model fit, i.e., a lower Akaike information criterion (AIC) score. There were 1,559 observations for German and 1,829 ones for English models.

3 Results and discussion

3.1 Descriptive results

The percentage of correct translations averaged for cognates is shown in Figure 2. It is clear from the plot that our German participants showed higher accuracy than English ones, with "Pred.Low" level surprisingly having the lowest accuracy. This contrasts with our expectation in stimulus design and implies the difference between native language comprehension (as in our stimulus validation) and intercomprehension (as in the free translation experiment) tangled with different inputs. That is, native language comprehension is multiple-choise-based, reading-related comprehension while intercomprehension is translation-based, listening-related comprehension across languages.

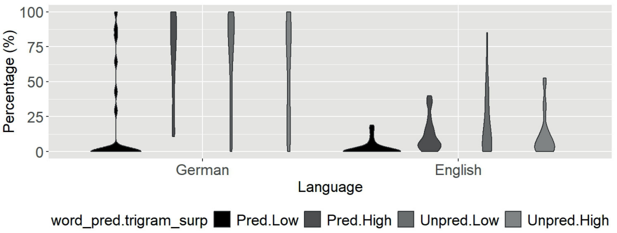


Figure 2: Percentage of correct translations for German and English averaged for items across levels of word predictability (word_pred being Pred or Unpred) and trigram surprisal (trigram_surp being Low or High). For each language, the violin diagrams are organized from left to right in the order of left to right in the legend.

3.2 GLMM results

Before implementing the GLMM models, we applied Variance Inflation Factor (VIF) and partial correlations to check the collinearity between the eight estimates. The VIF[8] reported values between 1 and 2 (*mean* = 1.39, *SD* = 0.21), suggesting a weak collinearity. We also found correlations among the estimates regardless of approaches.

To reduce the collinearity between the estimates and experimental factors, we extracted the residuals by implementing $lm(estimateX \sim trigram\_surp + word\_pred)$ and used the residuals instead of the estimates themselves in the GLMMs. We iteratively used one estimate and selected the best estimates from the two approaches by comparing AIC. While the best estimate of the cascaded approach was cognate surprisal (asrlm_cgnt_surp_resid), those of the LM-only were sentence entropy (lm_sent_entr_resid) for German and sentence surprisal for English (lm_sent_surp_resid).

The results of fixed effects are shown in Table 3. For both our German and English responses, we observed significant effects of those extracted from the LM-only approach but with different directions. Specifically, for our German participants, higher sentence entropy leads to lower accuracy (*Estimate* = -8.8947), meaning that the higher the uncertainty of the sentence-final cognates given their context, the participants are more likely translate the cognates incorrectly. This implies our German participants were influenced by the contextual cues. Whereas higher sentence surprisal leads to higher accuracy for our English participants (*Estimate* = 2.6463). This indicates that the higher the unexpectedness of our cognates given their context are, the higher the translation accuracy. This seems to suggest that when the contextual cues are not able to provide guidance, participants are less likely influenced by the context and focus only on the cognates themselves.

The non-flat standard errors for German could be due to the random slopes removing which leads to more stable estimates but similar results. On the other hand, providing additional written forms of the context led to significantly higher accuracy (*Estimate* = 0.5890), suggesting that our English participants encountered more difficulties compared to our German participants in spoken intercomprehension and leverage more direct information (i.e., written form rather than the context in spoken utterance itself).

Further, the estimates derived from the cascaded ASR-LM approach did not show significant contributions but

[5] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english/

[6] https://huggingface.co/benjamin/gerpt2/

[7] https://huggingface.co/openai-community/gpt2/

[8] via glm($is\_correct \sim all\_estimates$)

Language	Fixed Effect	Estimate	Std. Error	z value	Pr($>\|z\|$)	Sig.
German	(Intercept)	-3.6771	1.2473	-2.948	0.003197	**
	High	4.5669	1.3308	3.432	0.000600	***
	Unpred	5.3048	1.4812	3.581	0.000342	***
	AudioText	-0.2703	0.2349	-1.150	0.249941	
	asrlm_cognate_surprisal_resid	-3.8585	2.1252	-1.816	0.069432	
	lm_sentence_entropy_resid	-8.8947	2.4879	-3.575	0.000350	***
	High:Unpred	-5.6716	0.7376	-7.689	1.48e-14	***
English	(Intercept)	-4.7178	0.5049	-9.344	$<$ 2e-16	***
	High	1.7842	0.3756	4.751	2.03e-06	***
	Unpred	2.5041	0.3616	6.925	4.36e-12	***
	AudioText	0.5890	0.1949	3.021	0.00252	**
	asrlm_cognate_surprisal_resid	1.2444	0.6743	1.846	0.06496	
	lm_sentence_surprisals_resid	2.6463	0.5581	4.741	2.12e-06	***
	High:Unpred	-2.2191	0.4535	-4.893	9.91e-07	***

Table 3: Fixed effects of *glmer* models for German and English data. Significance (Sig.) codes: *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

fell in the same direction as those from the LM-only approach for each language. As the ASR models were "listening" to a foreign language and they may insert noises in their transcriptions, leading to insignificant contributions.

The contrasting results between German and English suggest that higher-level information (e.g., surprisal and entropy estimates related to predictability) is effectively utilized in comprehension when the two languages have closer typological proximity [30], as seen with Dutch and German. In contrast, when the linguistic distance is greater, as with English, listeners struggle to leverage contextual cues as they would in their native language.

In addition, we found significant main effects of trigram_surp (High) and word_pred (Unpred), both of which positively affected translation accuracy, as also shown in Figure 2. This outcome contradicts our initial expectations from stimulus validation, which was that sentences less preferred by Dutch native speakers and trigrams with high surprisal values would hinder cognate recognition. These findings suggest that cross-language comprehension does not necessarily follow the same patterns as native language comprehension. However, in trigram_surp, we observed a significant interaction with word_pred, indicating that shifting from predictable to unpredictable sentences reduces the likelihood of correct translations, which is consistent with our expectation in stimulus validation.

3.3 Limitations

Our GLMM results did not fully align with our expectations in stimulus validation. However, within High surprisal trigrams, lower word predictability (Unpred) was associated with reduced translation accuracy. Notably, our surprisal manipulation was confounded with syntactic structure: high-surprisal trigrams were prepositional phrases and low-surprisal ones were noun phrases. This confound issue may have influenced the unexpected outcomes, highlighting the need for future work to separate the effects of surprisal from syntactic structure. Another limitation is that our English participants showed the opposite performance to our German participants. Future studies could explore the effects of their L2 languages, include comprehension questions, analyze word similarity (e.g., the degree of cognate overlap or linguistic distances) in context, or use stimuli validated across all three languages to better understand its role in intercomprehension.

4 Conclusion

We investigated how language models (LM)-based estimates (surprisal and entropy) relate to the translation accuracy of cognates when we manipulated the contextual predictability of target cognate words in sentences. We conducted two versions of the free translation experiment with (1) audio-only input and (2) context in written form alongside the audio input. We included three Germanic languages: Dutch, German, and English. These three languages have different distances: Dutch and German have generally lower language distances (i.e., higher typological proximity [31]) than Dutch and English concerning syntax (word order), grammar complexity, vocabulary (lexicons), etc [32]. This provided a rich, varied linguistic domain for studying. The surprisal and entropy estimates were extracted via two approaches: Dutch spoken utterances transcribed with an Automatic Speech Recognition (ASR) system and then scored with an LM, and Dutch sentences scored by an LM only. These two approaches aimed to emulate German and English speakers listening to Dutch spoken utterances and reading Dutch sentences, respectively.

Though our findings reveal a contrast between German and English participants in translating Dutch words, LM-based surprisal/entropy estimates generally can be used to explain our participants' responses to a certain extent. For German speakers, higher surprisal and entropy values were associated with decreased translation accuracy, whereas English speakers exhibited the opposite pattern. This suggests that when two languages share greater typological similarity (e.g., Dutch vs. German), listeners can leverage contextual predictability, much like in native language processing. In contrast, speakers of more distant languages (e.g., English) struggle to make use of contextual cues. Additionally, the expected effects of sentence- and phrase-level predictability did not align with our predictions based on our stimulus validation. This discrepancy highlights a fundamental difference between cross-language comprehension and monolingual (native) processing.

5 Acknowledgements

This research is supported by a Rubicon grant from the Dutch Research Council to author WX (019.223-SG.004) and a grant by German Research Foundation (Project ID 232722074-SFB1102) to authors DK and BM.

References

[1] D. Doyé, *Intercomprehension*. Council of Europe Publishing, 2005.

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

[3] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?," *Language, Cognition and Neuroscience*, vol. 31, no. 1, pp. 32–59, 2016.

[4] F. Huettig, "Four central questions about prediction in language processing," *Brain Research*, vol. 1626, pp. 118–135, 2015.

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

[6] J. Aydelott and E. Bates, "Effects of acoustic distortion and semantic context on lexical access," *Language and Cognitive Processes*, vol. 19, pp. 29–56, 2004.

[7] J. Aydelott, D. Baer-Henney, M. Trzaskowski, R. Leech, and F. Dick, "Sentence comprehension in competing speech: Dichotic sentence word priming reveals hemispheric differences in auditory semantic processing," *Language and Cognitive Processes*, vol. 27, no. 7–8, pp. 1108–1144, 2012.

[8] C. D. Martin, G. Thierry, J.-R. Kuipers, B. Boutonnet, A. Foucart, and A. Costa, "Bilinguals reading in their second language do not predict upcoming words as native readers do," *Journal of memory and language*, vol. 69, no. 4, pp. 574–588, 2013.

[9] A. G. de Varda, M. Marelli, and S. Amenta, "Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data," *Behavior Research Methods*, pp. 1–24, 2023.

[10] I. Zaitova, I. Stenger, M. U. Butt, and T. Avgustinova, "Cross-linguistic processing of non-compositional expressions in Slavic languages," in *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024* (M. Zock, E. Chersoni, Y.-Y. Hsu, and S. de Deyne, eds.), (Torino, Italia), pp. 86–97, ELRA and ICCL, May 2024.

[11] 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.

[12] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "CELEX2 LDC96L14 [Web Download]," 1995.

[13] E. Keuleers, M. Brysbaert, and B. New, "SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles, journal = Behavior Research Methods," no. 3, 2010.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] W. de Vries and M. Nissim, "As good as new. How to successfully recycle English GPT-2 to make models for other languages," 2020.

[20] P. Boersma, "Praat: A system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[21] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: A unified javascript framework for online studies," in *International conference on computational social science (cologne)*, pp. 1–3, University of Osnabrück Cologne, 2017.

[22] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in English." `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`, 2021.

[23] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proceedings of Interspeech 2021*, pp. 2426–2430, 2021.

[24] B. Minixhofer, "GerPT2: German large and small versions of GPT2," 12 2020.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[26] K. Misra, "Minicons: Enabling flexible behavioral and representational analyses of transformer language models," *arXiv preprint arXiv:2203.13112*, 2022.

[27] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

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

[29] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[30] C. Gooskens and V. van Heuven, *Mutual Intelligibility*, pp. 51–95. Studies in Natural Language Processing, Cambridge University Press, 2021.

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

[32] N. Radman, L. Jost, S. Dorood, C. Mancini, and J.-M. Annoni, "Language distance modulates cognitive control in bilinguals," *Scientific Reports*, vol. 11, no. 1, p. 24131, 2021.