Simulating bilingual reading with language models: Effects of word frequency on cognate and interlingual homograph processing

Iza Škrjanec[1], Irene Elisabeth Winther[2], Vera Demberg[1], Stefan L. Frank[2]
[1]Saarland University; [2]Radboud University
skrjanec@lst.uni-saarland.de

Background. Words from different languages can have identical orthography. *Cognates* share both form and meaning between languages (e.g., *bed* in English and Dutch) while *interlingual homographs* (IHs) share form but not meaning (e.g., *stem*, meaning 'voice' in Dutch). Behavioral studies find that cognates facilitate processing in bilingual speakers [1, 2] compared to non-cognate control words and to monolinguals. This might be due to frequency effects: because cognates appear in both languages, bilinguals encounter them more frequently than non-cognate words. In contrast, IHs in sentence context result in inhibition that surfaces as longer reading times [3, 4]. We use surprisal (Eq. 1) from language models (LMs) as a computational correlate to human processing effort to simulate bilingual reading. We train Dutch-English LMs and test them on cognate and IH items to find if there is a facilitation (lower surprisal for cognates than controls) / inhibition (higher surprisal for IH) effect. We additionally test if the bilingual LM separates languages by looking at frequency patterns. Results from [5] suggest that processing is non-selective.

Method. We train monolingual English and Dutch as well as bilingual Dutch-English LSTMs on transcribed speech (subtitles) as well as non-fiction (Wikipedia) and fiction. The monolingual LMs were each trained on 10 million sentences, with a vocabulary of 32k words each. The bilingual LM was set up with budgeting for each language: Dutch as L1 (7.5 million sentences) and English as L2 (2.5 million sentences) with a joint vocabulary (24k Dutch, and 8k English words). The sentences were presented in random order. The bilingual LM simulates simultaneous unbalanced bilinguals. We test LMs on sentence-level stimuli. Dutch-English cognate words appear in English sentences [2], while IH words are embedded in Dutch [6]. We calculate the surprisal of target words and then fit linear mixed-effects regression models with surprisal as the response variable.

Results. Neither the bilingual nor monolingual models show an effect of word type (either for cognates or IH) when word type is the only fixed effect. Upon including language unspecific word frequency as seen in the training data, the bilingual models show a main effect of word type and of frequency (Table 1). The bilingual regression models with split Dutch and English frequency reveal that the target-language frequency predicts surprisal, while the word type and the non-target frequency are not significant (Table 2).

Discussion. The bilingual LM shows language-specific frequency effects: only the target language frequency modulates the estimated processing cost. This holds for cognates (where, in our simulations, the target language is L2) and IHs (with L1 as the target). This suggests that both L1 and L2 word frequency should be considered in analyses of cognate/IH processing in sentences. The regressions with joint frequency in bilingual models show an apparent positive word type (inhibition for cognates and IHs) and negative frequency effect. We explain this as follows: joint frequency is a significant predictor, but it overestimates single-language frequency (as only the target-language portion is informative), so the effect for word type corrects for this. Previous work [5] found facilitation effects for cognates, while we find inhibition. Our implementation differs in model size (fewer parameters) and dataset design (a 10× larger and more genre-diverse training set). We will follow their setup to replicate their work. Future experiments will include the French-English pair as well.

		Bilingual				Monolingual			
	Predictor	β	*SE*	t	p	β	*SE*	t	p
Cognates	(Intercept)	-2.19	1.02	-2.15	<.05	1.16	1.25	0.92	>.05
	Word type	0.54	0.08	6.55	<.001	-0.009	0.07	-0.13	>.05
	Frequency	-2.31	0.22	-10.46	<.001	-1.61	0.29	-5.61	<.001
IH	(Intercept)	3.53	1.46	2.42	<.05	-4.52	1.49	-3.04	<.01
	Word type	0.26	0.09	2.84	<.01	0.1	0.07	1.37	>.05
	Frequency	-1.15	0.31	-3.77	<.001	-2.83	0.32	-8.98	<.001

Table 1: Word type coding: -1 control, 1 cognate/IH. Bilingual LMs use joint frequency.

	Predictor	β	*SE*	t	p
Cognates	(Intercept)	-1.8	1.42	-1.27	>.05
	Word type	-0.09	0.29	-0.32	>.05
	NL frequency (non-target)	0.09	0.31	0.3	>.05
	EN frequency (target)	-2.47	0.43	-5.81	<.001
IH	(Intercept)	-2.35	2.22	-1.06	>.05
	Word type	-0.12	0.34	-0.34	>.05
	NL frequency (target)	-2.69	0.36	-7.58	<.001
	EN frequency (non-target)	0.23	0.28	0.81	>.05

Table 2: Results for bilingual LMs with language-specific frequency predictors from Dutch (NL) and English (EN) training data.

Examples of the cognate (1) and interlingual homograph (2) stimuli:

1. He convinces her to buy the **bed**$_{cognate}$/**art**$_{control}$
2. Anna rilde toen ze een **stem**$_{IH}$/**kerk**$_{control}$ opmerkte

$$surprisal(word) = -log\ p(word|context) \tag{1}$$

References

[1] Dijkstra, T., Grainger, J., & van Heuven, W. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *JML*.

[2] Bultena, S., Dijkstra, T., & van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Q. J. Exp. Psychol.*

[3] Libben, M., & Titone, D. A. (2009). Bilingual lexical access in context: Evidence from eye movements during reading. *J. Exp. Psychol.: Learn. Mem. Cogn.*

[4] Pivneva, I., Mercier, J., & Titone, D. A. (2014). Executive control modulates cross-language lexical activation during L2 reading: Evidence from eye movements. *J. Exp. Psychol.: Learn. Mem. Cogn.*

[5] Winther, I. E., Matusevych, Y., & Pickering, M. J. (2021). Cumulative frequency can explain cognate facilitation in language models. *Proceedings of CogSci*.

[6] Škrjanec, I., Winther, I. E., Huisman, M., Demberg, V., Bultena, S., & Frank, S. L. (2025). Slower reading on interlingual homographs can be a surprisal effect. *Book of abstracts AMLaP*.