A Model Information

Table 5 offers a model description in terms of size and languages, HuggingFace names, as well as paper references.

Model	Reference	Layers	Lang.	Size
google-bert/bert-base-uncased	Devlin et al., 2019	12	EN	110M
google-bert/bert-large-uncased	Devlin et al., 2019	24	EN	330M
FacebookAI/roberta-base	Liu et al., 2019	12	EN	125M
FacebookAI/roberta-large	Liu et al., 2019	24	EN	355M
almanach/camembertv2-base	Antoun et al., 2024	12	FR	112M
FlauBERT/flaubert_base_cased	Le et al., 2020	12	FR	138M
FlauBERT/flaubert_large_cased	Le et al., 2020	24	FR	373M
google-bert/bert-base-multilingual-uncased	Devlin et al., 2019	12	multi	179M
FacebookAI/xlm-roberta-base	Conneau et al., 2019	12	multi	279M
FacebookAI/xlm-roberta-large	Conneau et al., 2019	24	multi	560M
meta-llama/Llama-3.1-8B	Grattafiori et al., 2024	32	multi	8B
meta-llama/Llama-3.2-3B	Grattafiori et al., 2024	28	multi	3B

Table 5: Models used in this experiment can be obtained from https://huggingface.co.

B GPU Use

We ran the experiments with Llama 3.1 8B and LLama 3.2 3B on an Nvidia Ampere A100 80GB GPU node. We estimate the overall execution time to be 200 hours.

C Sentence-wise Surprisal

Sentence surprisal for EN models, word acceptability			Sentence surprisal for FR models, word acceptability			Sent. surprisal for FR models, pattern acceptability	
Model	**Default**	**K & I**	**Model**	**Default**	**K & I**	**Default**	**K & I**
$\text{BERT}_{\text{BASE}}$	-0.46*	-0.54*	$\text{CamemBERTv2}_{\text{BASE}}$	0.02	-0.06	-0.2	-0.11
$\text{BERT}_{\text{LARGE}}$	-0.46*	-0.55*	$\text{FlauBERT}_{\text{BASE}}$	-0.18	-	-0.11	-
$\text{RoBERTa}_{\text{BASE}}$	-0.57*	**-0.63***	$\text{FlauBERT}_{\text{LARGE}}$	-0.17	-	-0.06	-
$\text{RoBERTa}_{\text{LARGE}}$	-0.54*	-0.61*	$\text{XLM-RoBERTa}_{\text{BASE}}$	-0.24	-0.15	-0.3	0.01
$\text{XLM-RoBERTa}_{\text{BASE}}$	-0.36*	-0.54*	$\text{XLM-RoBERTa}_{\text{LARGE}}$	**-0.32***	-0.23	**-0.52***	-0.21
$\text{XLM-RoBERTa}_{\text{LARGE}}$	-0.42*	-0.65*	$\text{mBERT}_{\text{BASE}}$	0.02	0.02	-0.12	-0.02
$\text{mBERT}_{\text{BASE}}$	-0.16	-0.41*	Llama 3.2 3B	-0.2	-	-0.23	-
Llama 3.2 3B	-0.62*	-	Llama 3.1 8B	-0.19	-	-0.27	-
Llama 3.1 8B	-0.62*	-					

Table 6: Spearman correlaiton coefficients for sentence surprisal and human acceptability scores in English and French. For French, we additionally include the correlation with human acceptability as per polysemy pattern (averaged over acceptability of each word instance of a pattern). Asterisk (*) indicates a statistically significant result at $p < .05$.

D Semantic Similarity vs. Human Acceptability Judgment

We show mixed results obtained from the experiments with semantic similarity for selected models in Tables 7 to 10. Results are presented for RoBERTa, FlauBERT and Llama 3.1 8B covering both languages.

L.	Cosine	Cosine norm.	Spearman
RoBERTa$_{\text{BASE}}$			
0	-0.2, p=.034	0.08, p=.422	0.0, p=.964
1	-0.18, p=.058	0.27, p=.003	0.2, p=.031
2	-0.27, p=.003	0.37, p<.001	0.25, p=.006
3	-0.28, p=.002	0.44, p<.001	0.34, p<.001
4	-0.31, p<.001	**0.56, p<.001**	0.38, p<.001
5	-0.32, p<.001	0.54, p<.001	**0.42, p<.001**
6	-0.39, p<.001	0.42, p<.001	0.32, p<.001
7	-0.52, p<.001	0.25, p=.008	0.15, p=.105
8	-0.55, p<.001	0.1, p=.301	0.06, p=.494
9	**-0.58, p<.001**	-0.14, p=.129	-0.14, p=.144
10	-0.38, p<.001	-0.26, p=.006	-0.27, p=.003
11	-0.15, p=.102	-0.33, p<.001	-0.26, p=.005
12	-0.19, p=.04	-0.3, p=.001	-0.23, p=.013
RoBERTa$_{\text{LARGE}}$			
0	-0.02, p=.808	-0.0, p=.959	-0.05, p=.6
1	-0.15, p=.121	0.15, p=.104	0.0, p=.966
2	-0.32, p<.001	0.27, p=.004	-0.07, p=.485
3	-0.36, p<.001	0.29, p=.001	0.0, p=.974
4	-0.29, p=.002	0.32, p<.001	0.22, p=.017
5	-0.14, p=.131	0.36, p<.001	0.17, p=.062
6	-0.16, p=.086	0.43, p<.001	0.2, p=.034
7	-0.21, p=.022	0.48, p<.001	0.33, p<.001
8	-0.2, p=.029	**0.5, p<.001**	0.31, p<.001
9	-0.26, p=.005	0.47, p<.001	0.29, p=.002
10	-0.17, p=.078	0.41, p<.001	0.16, p=.083
11	0.01, p=.923	0.36, p<.001	0.09, p=.325
12	-0.25, p=.007	0.42, p<.001	0.14, p=.123
13	-0.17, p=.078	0.32, p<.001	0.06, p=.508
14	-0.21, p=.022	0.18, p=.048	-0.0, p=.99
15	-0.23, p=.016	0.11, p=.238	-0.09, p=.316
16	-0.23, p=.014	0.13, p=.166	-0.06, p=.549
17	-0.21, p=.023	0.11, p=.254	-0.08, p=.376
18	-0.2, p=.029	-0.05, p=.589	-0.2, p=.033
19	-0.37, p<.001	-0.18, p=.059	-0.26, p=.006
20	**-0.48, p<.001**	-0.29, p=.002	-0.18, p=.057
21	-0.48, p<.001	-0.36, p<.001	-0.12, p=.188
22	-0.34, p<.001	-0.46, p<.001	-0.2, p=.028
23	-0.0, p=.974	-0.49, p<.001	**-0.34, p<.001**
24	-0.4, p<.001	-0.26, p=.005	-0.21, p=.025

Table 7: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$ models.

L.	Cosine	Cosine norm.	Spearman
FlauBERT$_{\text{BASE}}$			
0	-0.01, p=.937	0.15, p=.356	-0.05, p=.764
1	0.05, p=.763	0.19, p=.22	0.08, p=.623
2	0.03, p=.848	0.19, p=.221	0.05, p=.773
3	0.02, p=.901	0.22, p=.17	0.04, p=.796
4	0.02, p=.882	0.24, p=.126	0.07, p=.677
5	0.04, p=.825	0.24, p=.125	0.07, p=.673
6	0.01, p=.935	0.21, p=.187	0.05, p=.737
7	0.05, p=.736	0.2, p=.205	0.04, p=.793
8	0.06, p=.696	0.21, p=.181	0.1, p=.51
9	0.11, p=.491	0.15, p=.331	0.09, p=.562
10	0.09, p=.55	0.15, p=.347	0.09, p=.578
11	0.13, p=.404	0.21, p=.182	0.13, p=.401
12	**0.31, p=.049**	**0.35, p=.023**	**0.33, p=.031**
FlauBERT$_{\text{LARGE}}$			
0	-0.05, p=.765	0.17, p=.293	-0.06, p=.719
1	-0.08, p=.628	0.18, p=.241	0.04, p=.81
2	-0.11, p=.471	0.14, p=.375	-0.07, p=.647
3	-0.15, p=.352	0.09, p=.553	-0.05, p=.76
4	-0.32, p=.041	0.04, p=.79	-0.15, p=.359
5	-0.34, p=.026	0.1, p=.513	-0.12, p=.464
6	-0.35, p=.022	0.2, p=.207	-0.08, p=.634
7	-0.31, p=.045	0.4, p=.01	-0.01, p=.944
8	-0.32, p=.036	0.45, p=.003	0.09, p=.577
9	-0.41, p=.008	0.45, p=.002	0.06, p=.684
10	-0.39, p=.011	0.45, p=.003	0.12, p=.436
11	-0.4, p=.009	0.45, p=.003	0.16, p=.316
12	-0.38, p=.014	0.46, p=.002	0.15, p=.329
13	-0.38, p=.014	**0.5, p<.001**	0.14, p=.369
14	**-0.42, p=.006**	0.42, p=.005	0.12, p=.43
15	-0.39, p=.01	0.37, p=.016	0.13, p=.42
16	-0.41, p=.007	0.38, p=.014	0.16, p=.324
17	-0.35, p=.023	0.42, p=.006	0.19, p=.232
18	-0.31, p=.044	0.43, p=.005	0.14, p=.38
19	-0.35, p=.023	0.41, p=.008	0.15, p=.332
20	-0.32, p=.036	0.39, p=.011	0.13, p=.414
21	-0.24, p=.121	0.38, p=.013	0.12, p=.438
22	-0.23, p=.136	0.39, p=.01	0.13, p=.395
23	-0.21, p=.18	0.4, p=.01	0.11, p=.489
24	-0.25, p=.106	0.44, p=.003	0.15, p=.336

Table 8: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by FlauBERT$_{\text{BASE}}$ and FlauBERT$_{\text{LARGE}}$.

Llama 3.1 8B (on English)			
L.	**Cosine**	**Cosine norm.**	**Spearman**
0	-0.01, p=.886	-0.16, p=.095	0.07, p=.464
1	0.02, p=.814	-0.01, p=.938	0.13, p=.164
2	-0.03, p=.749	0.19, p=.039	0.39, p<.001
3	-0.13, p=.157	0.34, p<.001	0.49, p<.001
4	0.06, p=.522	**0.54, p<.001**	**0.66, p<.001**
5	0.07, p=.449	0.5, p<.001	0.61, p<.001
6	0.13, p=.164	0.41, p<.001	0.51, p<.001
7	0.21, p=.026	0.43, p<.001	0.54, p<.001
8	0.28, p=.002	0.43, p<.001	0.54, p<.001
9	0.42, p<.001	0.4, p<.001	0.5, p<.001
10	0.42, p<.001	0.41, p<.001	0.52, p<.001
11	0.48, p<.001	0.4, p<.001	0.48, p<.001
12	0.51, p<.001	0.44, p<.001	0.51, p<.001
13	0.58, p<.001	0.48, p<.001	0.54, p<.001
14	**0.58, p<.001**	0.49, p<.001	0.54, p<.001
15	0.54, p<.001	0.48, p<.001	0.5, p<.001
16	0.52, p<.001	0.43, p<.001	0.47, p<.001
17	0.47, p<.001	0.42, p<.001	0.49, p<.001
18	0.38, p<.001	0.38, p<.001	0.45, p<.001
19	0.36, p<.001	0.38, p<.001	0.46, p<.001
20	0.32, p<.001	0.36, p<.001	0.47, p<.001
21	0.31, p<.001	0.35, p<.001	0.47, p<.001
22	0.29, p=.002	0.26, p=.005	0.41, p<.001
23	0.25, p=.007	0.19, p=.041	0.35, p<.001
24	0.22, p=.016	0.16, p=.08	0.33, p<.001
25	0.2, p=.03	0.13, p=.152	0.3, p=.001
26	0.19, p=.042	0.11, p=.222	0.29, p=.001
27	0.21, p=.022	0.1, p=.28	0.27, p=.003
28	0.15, p=.109	0.09, p=.328	0.29, p=.002
29	0.12, p=.194	0.07, p=.486	0.26, p=.005
30	0.11, p=.228	0.05, p=.563	0.25, p=.007
31	0.04, p=.648	0.06, p=.54	0.21, p=.023
32	-0.26, p=.004	0.07, p=.428	-0.2, p=.032

Table 9: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by Llama 3.1 8B, on English.

Llama 3.1 8B (on French)			
L.	**Cosine**	**Cosine norm.**	**Spearman**
0	-0.04, p=.812	0.12, p=.478	-0.02, p=.885
1	0.04, p=.816	0.05, p=.779	-0.01, p=.943
2	-0.13, p=.444	0.29, p=.076	0.03, p=.867
3	-0.23, p=.17	0.06, p=.74	-0.01, p=.947
4	-0.29, p=.073	0.11, p=.525	0.08, p=.651
5	-0.21, p=.214	0.31, p=.059	0.16, p=.348
6	-0.06, p=.741	0.4, p=.013	0.19, p=.263
7	0.09, p=.596	**0.41, p=.01**	0.2, p=.228
8	-0.06, p=.703	0.37, p=.022	0.2, p=.239
9	-0.03, p=.859	0.3, p=.07	0.19, p=.263
10	0.06, p=.723	0.27, p=.099	0.18, p=.27
11	0.17, p=.31	0.3, p=.066	0.2, p=.234
12	0.12, p=.455	0.26, p=.121	0.14, p=.388
13	0.06, p=.725	0.25, p=.128	0.2, p=.238
14	0.12, p=.492	0.24, p=.141	0.2, p=.226
15	0.09, p=.608	0.15, p=.385	0.12, p=.472
16	0.04, p=.805	0.09, p=.609	0.01, p=.94
17	-0.01, p=.968	0.12, p=.478	0.11, p=.518
18	-0.09, p=.591	0.1, p=.569	0.14, p=.411
19	-0.14, p=.392	0.09, p=.605	0.13, p=.431
20	-0.22, p=.185	0.1, p=.555	0.1, p=.55
21	-0.23, p=.156	0.09, p=.581	0.04, p=.811
22	-0.24, p=.14	0.1, p=.541	0.02, p=.895
23	-0.25, p=.133	0.07, p=.655	-0.0, p=.996
24	-0.27, p=.107	0.15, p=.364	0.0, p=.98
25	-0.27, p=.097	0.1, p=.537	0.04, p=.806
26	-0.3, p=.065	0.17, p=.316	0.07, p=.695
27	-0.24, p=.139	0.16, p=.345	0.09, p=.611
28	-0.26, p=.117	0.16, p=.34	0.1, p=.559
29	-0.26, p=.113	0.17, p=.316	0.11, p=.524
30	-0.33, p=.04	0.14, p=.393	0.02, p=.888
31	-0.29, p=.074	0.1, p=.549	-0.1, p=.541
32	**-0.33, p=.046**	0.09, p=.609	-0.21, p=.21

Table 10: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by Llama 3.1 8B, on French.

E Analysis of Model Scoring

Figure 4 demonstrates the surprisal score distribution in French for three types of senses as described in Section 3.1. The data on human evaluation is absent from this figure, since it was not obtained by the authors of the initial psycholinguistic experiment (Lombard et al., 2023).

Figure 5 shows surprisal score distribution in French as per regularity degree, along with human acceptability judgment.

Table 11 reveals the relationships of the models with the degree of pattern regularity, while Figure 4 shows correlation with word frequency and word length, as compared to the human evaluators.

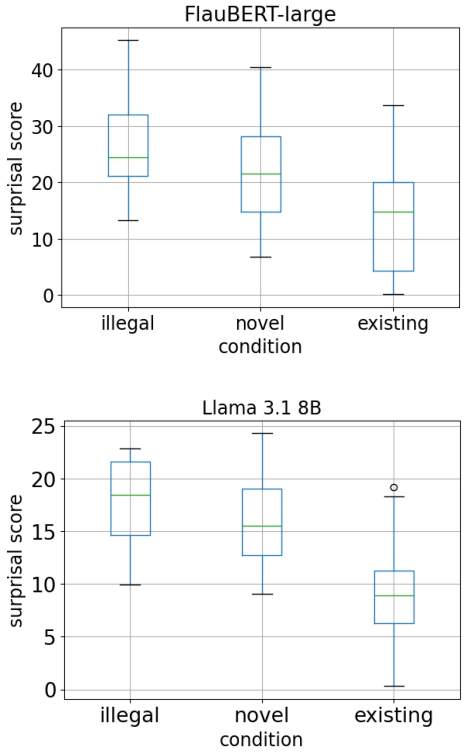

Figure 4: Distribution of French surprisal scores by condition labeled in the original dataset as *new*, *illegal* and *existing*. These correspond to the groups (1), (2) and (3) described in the Section 3.1

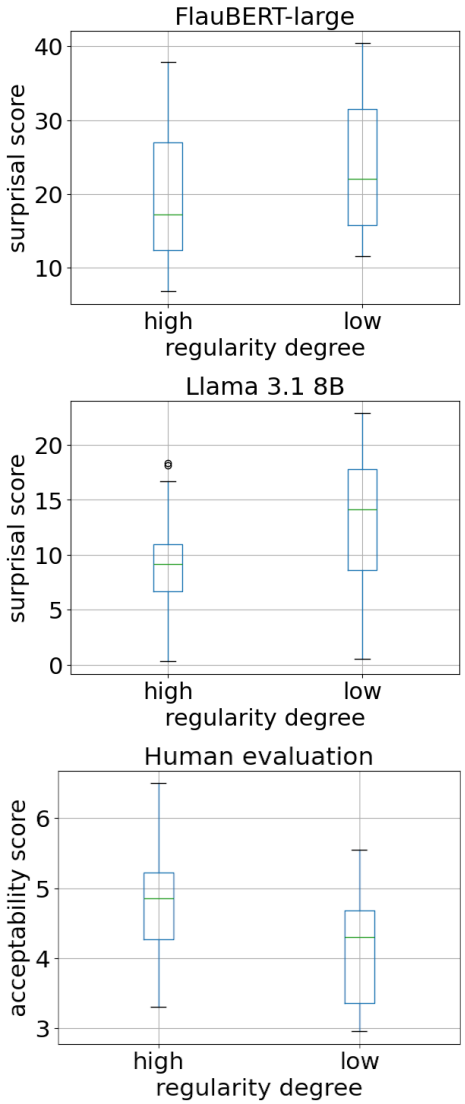

Figure 5: Distribution of surprisal scores by regularity degrees labeled as *high* and *low* for the French data.

Models	R1	R2	R3	R4
RoBERTa$_{\text{LARGE}}$ SURP	0.36*	0.43*	0.18	**0.44***
Llama 3.1 8B SURP	0.26*	**0.43***	0.01	0.3*
RoBERTa$_{\text{LARGE}}$ SIMIL	0.08	0.17	0.22	**0.27***
Llama 3.1 8B SIMIL	0.03	0.12	0.03	0.08
Human acceptability rating	0.28*	**0.53***	0.00	0.31*

Table 11: Correlation between the regularity metrics and scores from surprisal, similarity and human evaluation. The regularity metrics are defined in Table 1. Asterisk (*) indicates a statistically significant correlation ($p < .05$).

Limitations

In our experiments, we mainly focus on English data. The size of the dataset in French often did not allow us to make meaningful comparisons or confirm the validity of results received from English. As shown by Srinivasan and Rabagliati (2015), polysemy patterns and their degree of regularity overlap only partially across languages. Hence, studying the phenomenon of continuous regular polysemy in one language has limited generalization potential. More languages from diverse language families need to be involved in such investigations.

Regarding the models represented in our study, it would be interesting to compare masked BERT-based varieties with an autoregressive model that

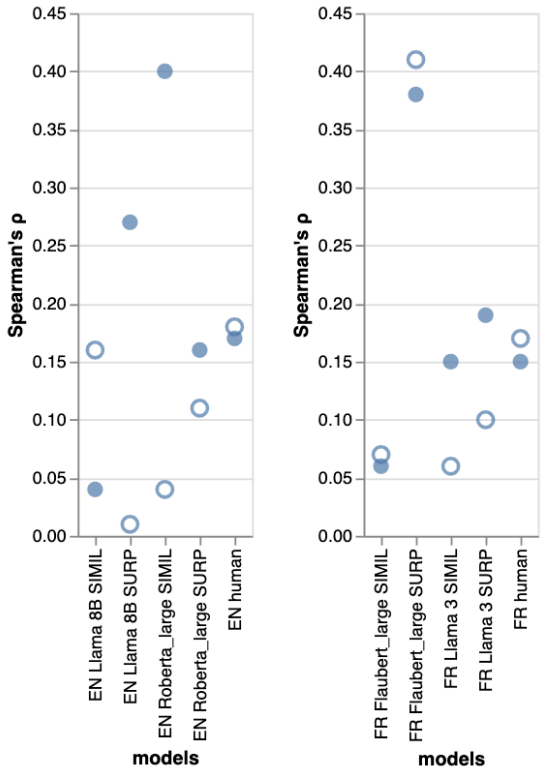

Figure 6: Correlation of word frequency and word length with the measures of semantic relatedness, surprisal and human rating. The empty circles represent word length, while the filled circles represent word frequency.

actually matches their size. Certain differences in polysemy processing were difficult to interpret since both varying architectures and wide-ranging model sizes were involved.