SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese

Nathan S. Hartmann – nathansh@icmc.usp.br
Gustavo H. Paetzold – g.h.paetzold@sheffield.ac.uk
Sandra M. Aluísio – sandra@icmc.usp.br

PROPOR 2018
September 24-26, 2018, Canela, Brazil

Introduction – What is Lexical Simplification?

Lexical Simplification (LS) has the function of changing words or expressions for synonyms that can be understood by a larger number of people (Paetzold and Specia, 2017b).

It is very common to have in mind a target audience, such as:

- Children
- People with cognitive disabilities
- Second language learners.

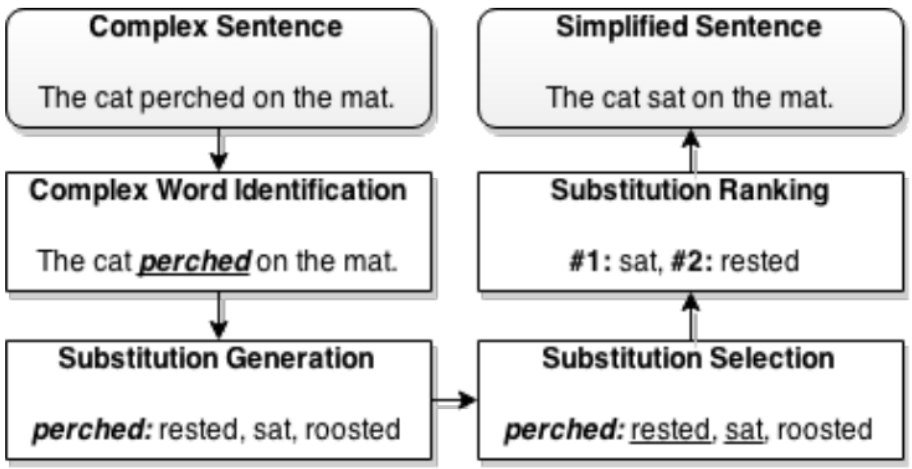

Figure 1: *Lexical simplification pipeline.*

Introduction – Background

In recent years there has been great activity in researching for:

- **English** (McCarthy and Navigli, 2007; De Belder and Moens, 2012; Specia et al., 2012; Yimam et al., 2017a, 2018)
- **Other languages** such as Japanese and **multilingual and cross-lingual** scenarios (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016; Yimam et al., 2017b, 2018).

Only two studies focus on children (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016).

There is no publicly available resource on LS for Portuguese.

Introduction – Contribution

This paper presents the SIMPLEX-PB:

- First publicly available corpus of LS for Brazilian Portuguese
  - Targeting children from the 1st to 9th school years.

- We also make available a benchmark
  - Evaluating the most well-known methods of LS.

SIMPLEX-PB – Target audience and complex words

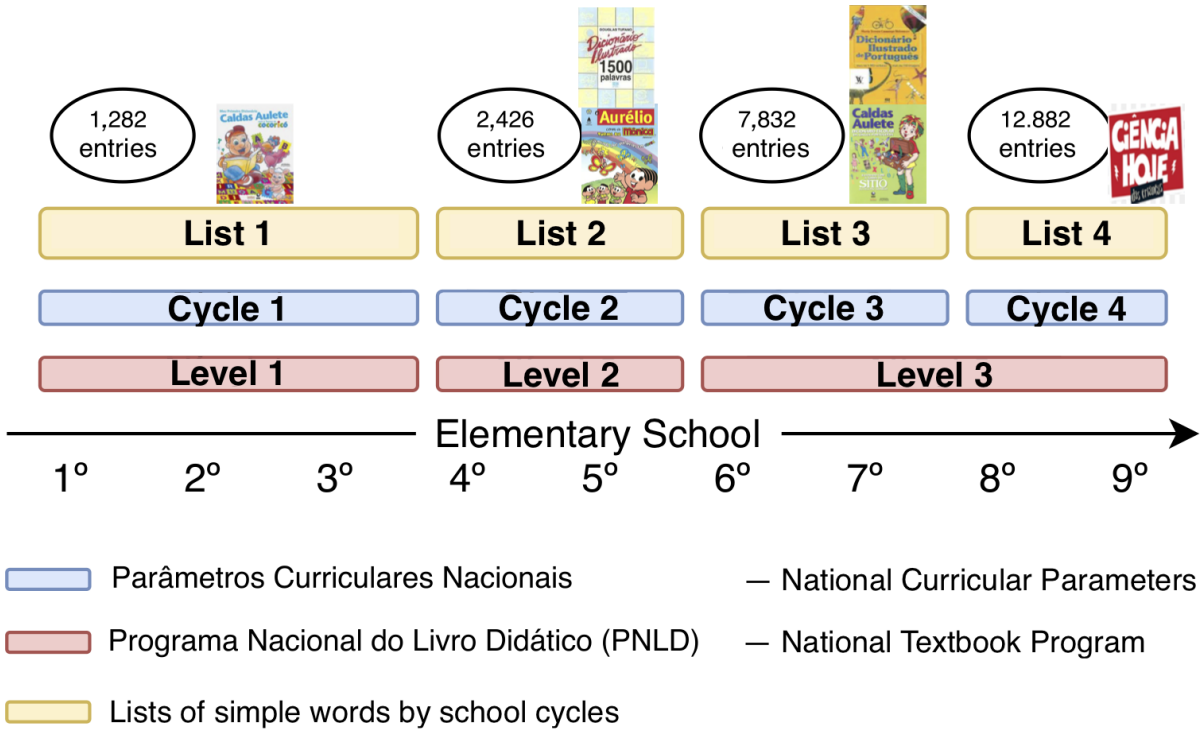

Figure 2: *Elementary School: cycles and complex words.*

Complex words are all those listed for the 8th and 9th years.

SIMPLEX-PB – Data acquisition

We used texts of several sources, written for children.

Textbooks	NILC corpus	SARESP tests	*Ciência Hoje das Crianças*	*Folhinha* Issue of Folha de São Paulo	*Para seu Filho Ler* Issue of Zero Hora	*Mundo Estranho*
492	262	72	2.589	308	166	3.756

Table 1: *Texts collected and annotated in Hartmann et al. (2016).*

16,170 sentences with exactly **one complex word** were identified.

SIMPLEX-PB – Corpus compilation

1,719 instances selected following the distribution of content words found in our corpus:

- 56% nouns
- 18% adjectives
- 18% verbs
- 6% adverbs.

Also made a subdivision equally distributed to include:

- Most frequent words
- Words with more synonyms
- Words with more senses.

757 distinct words were a target of simplification in our corpus.

SIMPLEX-PB – Corpus compilation

Given a complex word:

- Lemmatized it using DELAF (Muniz, 2004).
- Retrieved all its synonyms with same POS tag from TeP 2.0 (Maziero et al., 2008).
- Sorted them by their frequencies in a large corpus (Hartmann, 2016) and the appropriateness of the words to the given context:
  - Mean of cosine similarity between a synonym embedding and the context of the complex word (Hartmann, 2016).

SIMPLEX-PB – Annotation

- Substantial Cohen Kappa agreement of 0.73 (Cohen, 1960)
- An average of 1.43 ± 0.7 gold candidates per complex word
  - For English, LexMTurk (Horn et al., 2014) have an average of 12.64 ± 6.4 gold candidates and BenchLS (Paetzold and Specia, 2016) 7.36 ± 5.3.

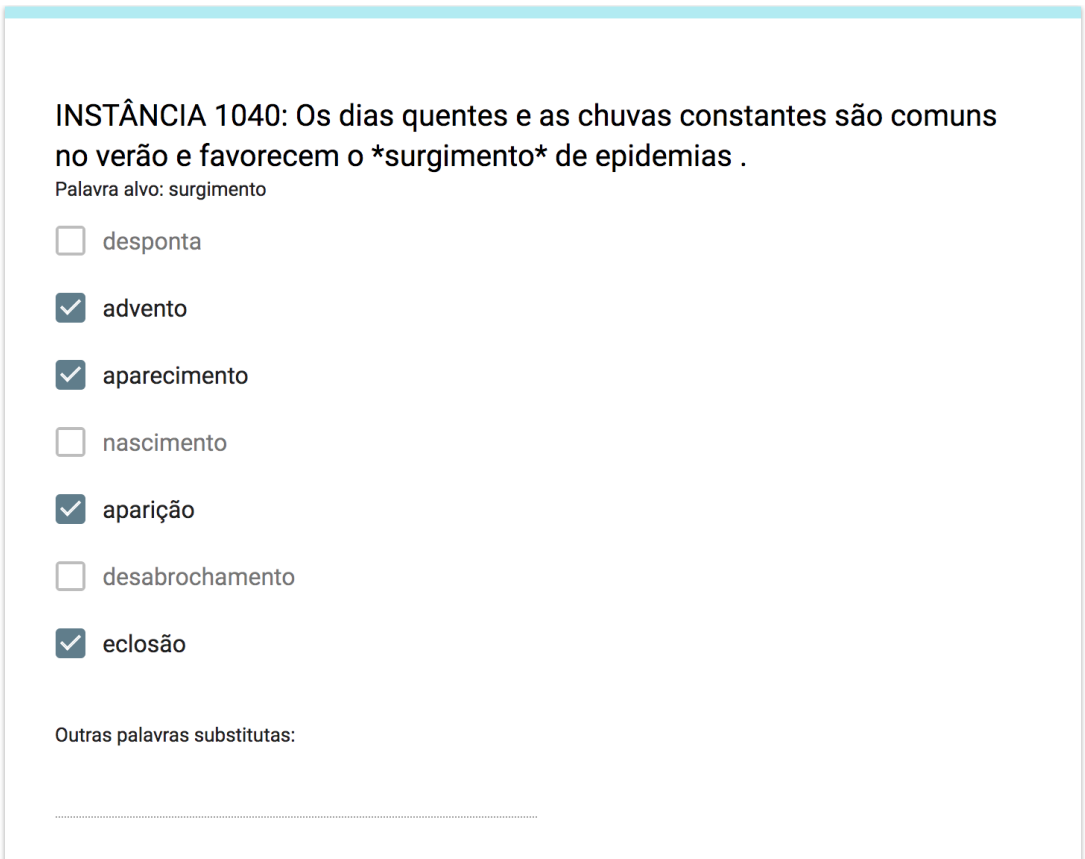

Figure 3: *Example of the annotation interface.*

Methods

Our LS methods for Brazilian Portuguese perform two steps:

- Substitution Generation
- Substitution Ranking.

Complex Word Identification is not performed.

Almost all LS contributions for English do the same.

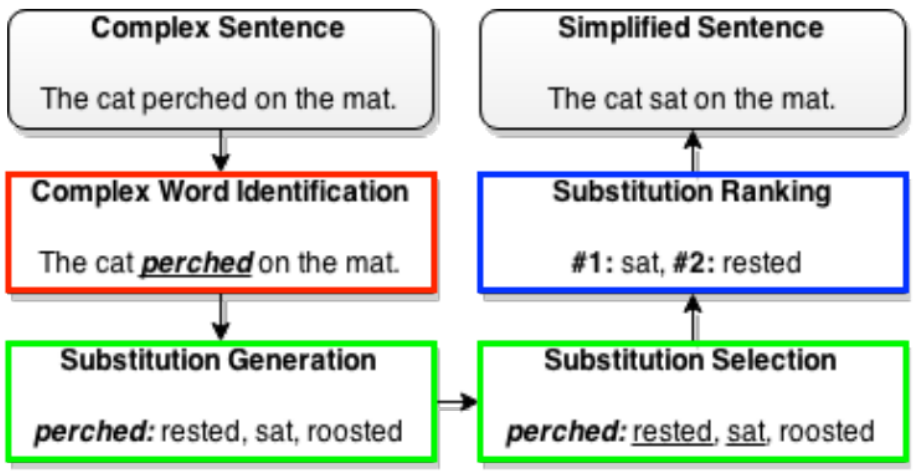

Figure 4: *The steps we address are colored in blue.*

Methods: Substitution Generation

- **Lexicon-Based:** Given a target complex word, synonyms are extracted from a lexicon
  - Synonyms taken from TeP 2.0 (Maziero et al., 2008).

- **Embeddings-Based:** Extracts the 10 non-morphological variants with the highest cosine similarity with the target word
  - 300d embedding models trained for Portuguese in a corpus of 1.4B tokens (Hartmann et al., 2017)
  - word2vec, wang2vec, glove and fasttext.

Evaluation: Substitution Generation Proficiency

- **Potential**: % of instances at least one gold candidate was generated
- **Precision**: % of generated words among the gold candidates
- **Recall**: % of gold candidates generated
- **F1**: harmonic mean between Precision and Recall.

	Potential	Precision	Recall	F1
TeP	**0.506**	**0.068**	**0.506**	**0.121**
glove	0.378	0.043	0.328	0.076
wang2vec	0.368	0.044	0.336	0.078
word2vec	0.335	0.038	0.291	0.068
fasttext	0.259	0.028	0.208	0.050

Table 2: *Candidate generation evaluation results.*

Methods: Substitution Ranking

- **Frequency-based (Frequency):** Ranks candidates according to their frequency in a large corpus
  - OpenSubtitles2016 corpus (Lison and Tiedemann, 2016).

- **Rank averaging (Rank Avg.):** Averages the rank across all features in order to produce a final ranking

- **Pairwise regression (Regression):** A ridge regression model that quantifies the simplicity difference between words
  - It receives features for a pair of candidate substitutions and the difference in simplicity between them.

Methods: Substitution Ranking

Rank Avg. and **Regression** approaches use the same 17 features:

- **Frequency of n-grams** surrounding the target complex word considering n-grams formed by $0 \leq n \leq 2$ tokens to its left and right (9 features)
- **Cosine similarity** between the candidate and the target complex word on the four embedding models (4 features)
- **Average cosine similarity** between the candidate and a window of 3 words surrounding the target on the four embedding models (4 features).

We chose these features for their effectiveness in English LS (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2017a).

Evaluation: Target Replacement Proficiency

- **Accuracy**: the proportion of instances for which the highest ranking candidate is within the gold simplifications.

We paired our Substitution Generators with our Substitution Rankers to produce 15 full simplifiers.

	Frequency	**Rank Avg.**	**Regression**
TEP	0.227	**0.292**	0.211
wang2vec	0.167	0.200	0.121
fasttext	0.172	0.175	0.118
glove	0.086	0.174	0.095
word2vec	0.112	0.158	0.094

Table 3: *Candidate replacement evaluation results.*

Conclusion

SIMPLEX

github.com/nathanshartmann/SIMPLEX-PB

Future work:

- Increase the list of gold candidates using this evaluation
- Propose new LS approaches
- Evaluate **complex word identification** methods
  - 2nd place in CWI shared-task of NAACL 2018 (Hartmann and Santos, 2018).

References I

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37–46.

De Belder, J. and M.-F. Moens (2012). A dataset for the evaluation of lexical simplification. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Berlin, Heidelberg, pp. 426–437. Springer Berlin Heidelberg.

Glavaš, G. and S. Štajner (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015)*, Volume 2, pp. 63–68.

Hartmann, N., L. Cucatto, D. Brants, and S. Aluísio (2016). Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco (Eds.), *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*, pp. 12–24. Springer International Publishing.

Hartmann, N., E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hartmann, N. S. (2016). ASSIN Shared Task - Solo Queue Group: Mix of a Traditional and an Emerging Approaches. In *Avaliação de Similaridade Semântica e Inferência Textual (ASSIN), Propor Workshop*.

Hartmann, N. S. and L. B. Santos (2018). NILC at CWI 2018: Exploring Feature Engineering and Feature Learning. In *Proceedings of BEA13*.

Horn, C., C. Manduca, and D. Kauchak (2014). Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the ACL (ACL-2014)*, pp. 458–463.

Kajiwara, T. and K. Yamamoto (2015). Evaluation dataset and system for japanese lexical simplification. In *ACL (Student Research Workshop)*, pp. 35–40. The Association for Computer Linguistics.

References II

Kodaira, T., T. Kajiwara, and M. Komachi (2016). Controlled and balanced dataset for japanese lexical simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pp. 1–7. Association for Computational Linguistics.

Lison, P. and J. Tiedemann (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th LREC*.

Maziero, E. G., T. A. Pardo, A. Di Felippo, and B. C. Dias-da Silva (2008). A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, pp. 390–392. ACM.

McCarthy, D. and R. Navigli (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 48–53. Association for Computational Linguistics.

Muniz, M. C. M. (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Universidade de São Paulo, Brasil.

Paetzold, G. and L. Specia (2017a). Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2*, pp. 34–40. ACL.

Paetzold, G. H. and L. Specia (2016). Benchmarking lexical simplification systems. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, 3074–3080.

Paetzold, G. H. and L. Specia (2017b). A survey on lexical simplication. *J. Artif. Intell. Res. 60*, 549–593.

Specia, L., S. K. Jauhar, and R. Mihalcea (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 1st SEM*, pp. 347–355. ACL.

Yimam, S. M., C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri (2018). A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th BEA*. Association for Computational Linguistics.

Yimam, S. M., S. Štajner, M. Riedl, and C. Biemann (2017a). Cwig3g2 - complex word identification task across three text genres and two user groups. In *Proceedings of the 8º IJCNLP*, pp. 401–407. Asian Federation of Natural Language Processing.

Yimam, S. M., S. Štajner, M. Riedl, and C. Biemann (2017b). Multilingual and cross-lingual complex word identification. In *Proceedings of RANLP*, pp. 813–822.

Thank you!!

Questions?