SMILLE for Portuguese: Annotation and analysis of grammatical structures in a pedagogical context

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon

Centre de traitement automatique du langage – CENTAL
Université catholique de Louvain (UCL), Belgium

Outline

Introduction

Noticing hypothesis

- Language learners have to notice the relevant information
  - "people learn about the things that they attend to and do not learn much about the things they do not attend to" [1].
- Authentic texts may not present properties for **drawing the learner's attention**

Input enhancements

- Aims to solve the lack of salience in raw input
- Enhancements of the learning (grammatical) content associated with SLA curriculum

Introduction

Related work

These are some works that provide different levels of input enhancements:

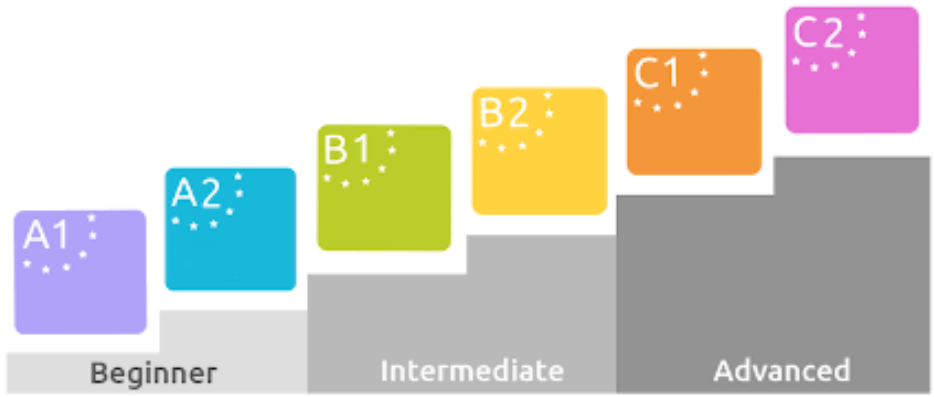
- REAP [2], WERTi [3], SmartReader [4, 5], FLAIR [6], and SMILLE [7, 8, 9]

Introduction

Related work

These are some works that provide different levels of input enhancements:

- REAP [2], WERTi [3], SmartReader [4, 5], FLAIR [6], and SMILLE [7, 8, 9]



Common **E**uropean **F**ramework of **R**eference for Languages [10]

Introduction

Smart and **I**mmersive **L**anguage **L**earning **E**nvironment

Introduction

Smart and **I**mmersive **L**anguage **L**earning **E**nvironment

SMILLE

- SMILLE applies **rules** on top of the **parser annotation**
  - PoS-tagger, dependency-parser, and infered information.
    E.g. hidden and explicit subject, and types of gerunds
- SMILLE links the detected information to the guidelines of Altissia International and the CEFR

Objective

1. To extend SMILLE for Portuguese
   - How reliable are the new rules?
2. To profile grammatical structures that occur in pedagogical texts, and to contrast them with structures presented in handbooks
   - How does the textual content presented to the learners align with the grammatical content that is taught in the different levels?

SMILLE for Portuguese

Rules

- 71 rules based on PassPort parser's output
- Brazilian and European variants
- CEFR levels from A1 to B2
- Each rule is linked to a specific level.

Grammatical structures

Prepositions, articles, use of pronouns "tu" and "você", pronouns used as indirect and direct complements, possessive pronouns, demonstrative pronouns, comparatives, adjectives, plural forms, nouns, expression of preferences, imperative, expressions of obligation, various verb tenses (including progressive ones), interrogative sentences, irregular verbs, uses of "ser", "estar", "ter" and "haver", diminutives, direct and indirect complements, superlative, final clauses, relative clauses and pronouns, verbal periphrases, numbers, possessives, indefinite pronouns, use of the pronoun "si", several types of adverbs and adverbs derived from adjectives, passive voice, hidden and explicit subjects, and use of clitics.

SMILLE for Portuguese

Evaluation

We used 1425 random sentences from 3 different genres and applied SMILLE's pipeline

1. Text genres:
   - Literature (from romances available at `www.dominiopublico.gov.br`)
   - Newspaper articles (from the Diário Gaúcho corpus)
   - Subtitles (from the Portuguese OpenSubtitles corpus [11]).
2. 25 sentences randomly extracted for each structure from each genre (totaling 75 sentences per structure)
3. Manual evaluation (1 judge)

SMILLE for PT evaluation

Evaluation (Precision)

- Most of the structures have high precision (mean = 84.07%; median = 88%)
- A few structures had a really bad precision in Newspaper articles, but most of them scored high. (mean = 82.32%; median = 89%).
- Subtitles had the best mean and median precision (mean = 86.32%; median = 96%)

SMILLE for Portuguese

SMILLE for PT: the evaluation of the grammatical structures

- The genre doesn't seem to be decisive for many grammatical structures
- Few grammatical structures presented precision biased by one genre *(e.g. compound future or the hidden or explicit pronominal subjects)*

Besides SMILLE's evaluation...

This precision evaluation showed us which **structures can be used** in further analyzing the pedagogical material in terms of content and organization per level.

Profiling pedagogical material

SMILLE aims to support reading activities...

- Can SMILLE be applied to observe the adequacy of texts used in Portuguese as foreign language courses?
  - How does SMILLE profile the texts of a course of Portuguese as foreign language in terms of grammar?

Profiling pedagogical material

SMILLE aims to support reading activities...

- Can SMILLE be applied to observe the adequacy of texts used in Portuguese as foreign language courses?
  - How does SMILLE profile the texts of a course of Portuguese as foreign language in terms of grammar?

Corpus

- Corpus of texts from the course of Brazilian Portuguese as foreign language of the UFJF

	Basic level	Intermediate level	Total
Tokens	8,031	11,710	19,741
Sentences	421	356	957

Profiling pedagogical material

SMILLE setup

- Curriculum of the specific course was used
- Only grammatical structures with precision above 80% were considered
  - We added 52 structures that are directly based on parser information *(E.g. simple verb tenses and word classes)*
    - Parser accuracy: 94% for PoS tagging and 85% for dependency parsing [12].

Steps

1. To study the corpus in terms of distribution of grammatical structures in each level
2. We contrasted the different prominences of grammatical structure with the handbooks' curriculum

Profiling pedagogical material

Significant results

- Some structures are significantly more prominent in the sentences of the basic level
  E.g. the present tense of the verb "ter" and the use of personal pronouns as subject
- At the intermediate level, the more prominent structures are those taught at the end of the basic level and are reviewed with more emphasis during the intermediate level
  E.g. the past future and the past imperfect tense

Profiling pedagogical material

Observations

- Most of the observed structures did not show a significant difference and occur in a similar way in both levels
- This type of analysis can aid teachers and pedagogical coordinators in the task of preparing a language course, so that the texts can better reflect and emphasize the focused grammatical content.

Final remarks

Our goals

1. To extend SMILLE for Portuguese
2. To evaluate the applicability of SMILLE in the identification of texts to be used in courses of Portuguese as foreign language

Final remarks

SMILLE for Portuguese

- 4 CEFR levels covered
- It can be used to
  - enhance texts that are to be used with language learners
  - help in text selection, supporting the teacher's activities

SMILLE for Portuguese: summary of the evaluation

- 3 different genres: newspaper articles, literature and subtitles
- Most of the structures scored as high as 100% of precision
- The system's average precision is 84%
- For the structures that presented bad performance:
  - we saw a mix of **bad parsing** performance and **bad rules**, so
  - we will be addressing these issues for the future versions of the system

Final remarks

To evaluate the applicability of SMILLE in the identification of texts to be used in courses of Portuguese as foreign language

- we could observe that the structures in the two levels generally follow the description provided in the handbooks for the basic and the intermediate levels.
- For some of the structures, there may be mismatched with the level's grammatical content

Final remarks

Future work

1. To expand the corpus of handbook texts
2. To include learners' texts
   - Then we can compare how both these instances of language learning behave in terms of grammar

References I


R. Schmidt, "Attention, awareness, and individual differences in language learning," *Perspectives on individual characteristics and foreign language education*, vol. 6, p. 27, 2012.


J. Brown and M. Eskenazi, "Retrieval of authentic documents for reader-specific lexical practice," in *InSTIL/ICALL Symposium 2004*, 2004.


D. Meurers, R. Ziai, L. Amaral, A. Boyd, A. Dimitrov, V. Metcalf, and N. Ott, "Enhancing authentic web pages for language learners," in *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 10–18, Association for Computational Linguistics, 2010.


M. Azab, A. Salama, K. Oflazer, H. Shima, J. Araki, and T. Mitamura, "An nlp-based reading tool for aiding non-native english readers," *Recent Advances in Natural Language Processing*, p. 41, 2013.


M. Azab, A. Salama, K. Oflazer, H. Shima, J. Araki, and T. Mitamura, "An english reading tool as a nlp showcase," in *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, (Nagoya, Japan), pp. 5–8, Asian Federation of Natural Language Processing, October 2013.


M. Chinkina, M. Kannan, and D. Meurers, "Online information retrieval for language learning," *ACL 2016*, p. 7, 2016.


L. Zilio and C. Fairon, "Adaptive system for language learning," in *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on*, pp. 47–49, IEEE, 2017.


L. Zilio, R. Wilkens, and C. Fairon, "Enhancing grammatical structures in web-based texts," in *Proceedings of the 25th EUROCALL*, pp. 839–846, Accepted, 2017.

References II


L. Zilio, R. Wilkens, and C. Fairon, "Using nlp for enhancing second language acquisition," in *Proceedings of Recent Advances in Natural Language Processing*, pp. 839–846, 2017.


N. Verhelst, P. Van Avermaet, S. Takala, N. Figueras, and B. North, *Common European Framework of Reference for Languages: learning, teaching, assessment*.
Cambridge University Press, 2009.


J. Tiedemann, "Finding alternative translations in a large corpus of movie subtitle.," in *International Conference on Language Resources and Evaluation*, 2016.


L. Zilio, R. Wilkens, and C. Fairon, "Passport: A dependency parsing model for portuguese," in *International Conference on Computational Processing of the Portuguese Language*, Springer, 2018.

Highlight Exercises

General Information

Level A1

- Adjectives (3)
- Determiners: Articles (1)
- Nouns: Plural (1)
- Nouns: Singular (23)
- Personal Pronouns (1)

Level A2

Level B1

Level B2

SMILLE for Portuguese: Annotation and analysis of grammatical structures in a pedagogical context

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon

Centre de traitement automatique du langage -- CENTAL

Université catholique de Louvain, Belgium

I may check out SMILLE in demo session today

SMILLE for Portuguese: Annotation and analysis of grammatical structures in a pedagogical context

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon

Centre de traitement automatique du langage -- CENTAL

Université catholique de Louvain, Belgium

Thank you!

I may check out SMILLE in demo session today