A web interface for language acquisition studies

Rodrigo Wilkens, Guilherme Schwade, and Aline Villavicencio

Federal University of Rio Grande do Sul
rodrigo.wilkens@inf.ufrgs.br
guilherme@wing.inf.br
avillavicencio@inf.ufrgs.br

1 Introduction

Child language databases are a valuable resource for language acquisition studies. They can provide the basis for developmental comparisons and evaluation of different hypotheses. For computational investigations annotated corpora can be used as an approximation to the linguistic environment to which a child is exposed, as discussed by Wintner in [1].

There are some initiatives to annotate data from children, with information about part-of-speech (PoS), syntactic and semantic aspects [5], WordNet [2] and MRC Psycholinguistic, and some of these are available as part of the English corpora in CHILDES [3]. However, for many other languages (like Portuguese) one of the challenges is to collect corpora and annotate it, as the resources and the tools are less abundant and many of them are still under development.

At present, CHILDES contains data for more than 25 languages and it is the most used source for data in computational language acquisition studies. An initial step, for Portuguese, was given by [7] which annotated the corpora in several dimensions (speech role, morphological, syntactic, semantic, psycholinguistic information) aiming to make it available as a database. In this work we address the problem of improving user access\(^1\) to the CHILDES Portuguese data in an user-friendly interface and providing in an easier way to perform precise searches and analysis both monolingually or cross-lingually, using the same pattern to search another language in the database.

2 Resource Description

The Portuguese CHILDES contains 3 corpora: (1) Batoréo, with 60 narratives, 30 from adults and 30 from children, about two stories (2) Porto Alegre with data from 5 to 9 year old children, collected both cross-sectionally and longitudinally; and (3) Florianópolis with the longitudinal data for one Brazilian child: 5530 utterances in broad phonetic transcription.

\(^1\) The system is available at aldebaran.lids.mit.edu/~rswilkens/guilherme/csdb
In contrast with English CHILDES the Portuguese data is much smaller\(^2\), but as presented previously by [7] it can be used to compare children development by age.

The corpora were normalized to deal with incomplete words and remove transcription annotations, and then automatically lemmatized, PoS tagged, parsed and assigned semantic tags (e.g. organization, date, place, etc.) with the PALAVRAS parser [4]. In addition the words were annotated with psycholinguistic information like age of acquisition and concreteness from [6]. And, finally, the corpus was stored in a database, a table with words (Tables 1b) and other to sentences (Tables 1a) [7].

### 3 System Description

To facilitate access to the CHILDES corpora, we developed a web interface that combine information from different levels of annotation. The data is organized in a database, structured as in Tables 1b and 1a, respectively presenting the structure of words and sentences.

<table>
<thead>
<tr>
<th>(a) Information about Sentences</th>
<th>(b) Information about Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Sentence</strong></td>
</tr>
<tr>
<td>age of acquisition</td>
<td>corpus information</td>
</tr>
<tr>
<td>part-of-speech</td>
<td>children gender</td>
</tr>
<tr>
<td>frequency by age from children</td>
<td>PoS tags</td>
</tr>
<tr>
<td>wordnet relations</td>
<td>dependency tree</td>
</tr>
<tr>
<td>adult frequency</td>
<td>semantic tags</td>
</tr>
</tbody>
</table>

Using this environment an user can choose any combination of fields in the database to perform a query (by key-word or regular expression search), and the result can be downloaded as a CSV file or can be displayed as a cumulative graph. Another important feature is the integration between languages (currently Portuguese and English), which enables running the same search on both languages. In this way it is possible to examine, for instance, the usage of a particular word and its evolution according to grammatical class per age and cross-lingually. Searches are done in two steps:

**Initial search:** with a text box to start the search process (inspired by Google search interface, so that the search process is close to other search experiences of the user).

\(^2\) The Portuguese CHILDES contains 23,078 sentences and the English one contains 4,845,264, in other words, the Portuguese CHILDES is 0.47% smaller than the English version.
Refine search allows the user to set filters to reduce and refine the returned results.

The system presented in this paper allows to create and reproduce, in an easy way, studies on language acquisition based on CHILDES. Some examples of possible analyses\(^3\) are:

- check the frequency of words, PoS and complements (direct, dative and prepositional) by age;
- identify if and when there is an inversion of the frequency of verbs and nouns across the ages;
- analyze the pro-drop phenomena by querying sentences with and without subject, and yet, compare a pro-drop language (Portuguese) with a non pro-drop language (English).

4 Conclusion

In this paper we presented an web interface for easier and integrated access to CHILDES database [7]. Our main goal with this system is to help language acquisition research by providing a tool able to perform complex queries in a simple way. This system is inspired by familiar search engines facilitating the learning curve, and it can be used in new research works.

References


\(^3\) The demonstration will present the four analyses proposed in this paper and others requested by audience.