I say have you say tem: profiling verbs in children data in English and Portuguese

Rodrigo Wilkens
Institute of Informatics
Federal University of Rio Grande do Sul
Brazil
rswilkens@inf.ufrgs.br

Aline Villavicencio
Institute of Informatics
Federal University of Rio Grande do Sul
Brazil
avillavicencio@inf.ufrgs.br

Abstract

In this paper we present a profile of verb usage across ages in child-produced sentences in English and Portuguese. We examine in particular lexical and syntactic characteristics of verbs and find common trends in these languages as children's ages increase, such as the prominence of general and polysemic verbs, as well as divergences such as the proportion of subject dropping. We also find a correlation between the age of acquisition and the number of complements of a verb for English.

1 Introduction

In this paper we report on a large scale investigation of some linguistic and distributional patterns of verbs in child-produced sentences for two languages, Portuguese and English. We compare the characteristics that emerge for two languages that, in spite of similarities in terms of verb usages also have important differences, in particular in allowing subject pro-drop, and examine to what degree these are reflected in the data. This is particularly relevant given the sparseness (and in some cases lack) of the Portuguese data, in particular for certain ages, which may not provide as clear indications as the English data, but existing analysis for the latter can also benefit the former and be used to help assess results obtained for similar trends found in it.

As such our work is related to that of Buttery and Korhonen (2007) who perform a large scale investigation of the subcategorization frames in the English corpora in CHILDES (MacWhinney, 2000), a database containing transcriptions of child-directed and child-produced sentences, comparing preferences in child and adult language to provide support for child language acquisition studies. These preferences are found using large amounts of automatically annotated data that would be otherwise too costly and time consuming to manually annotate.

At present, CHILDES contains data for more than 25 languages including English and Portuguese. For English, the corpora are currently available with annotations in raw, part-of-speech-tagged, lemmatized and parsed formats (Sagae et al., 2010) (Buttery and Korhonen, 2005) (Buttery and Korhonen, 2007). Although there are similar initiatives for other languages, like Spanish and Hebrew (Sagae et al., 2010), for Portuguese, there is a lack of such annotations on a large scale. In this work we address this issue and automatically annotate the Portuguese corpora with linguistic and distributional information using a robust statistical parser, providing the possibility of deeper analysis of language acquisition data.

Crosslinguistic investigations of child-produced language have also highlighted the important role of very general and frequent verbs, light verbs like *go, put* and *give* which are among the first to be acquired for languages like English and Italian as discussed by Goldberg (1999). In this paper we compare patterns found in child verb usage in English and Portuguese, in one of the first large scale investigations of syntactically annotated child-produced Portuguese data. Using this level of annotation we are able to examine patterns in verb usage in particular in terms of subjects and complements. Thus, this work is also related to the that of Valian (1991) who found a subject pro-drop rate of around 70% for 2 to 3 year old children in Italian, a pro-drop language,

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 38–42,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

and even a significant number of subject omission for English, which is not a pro-drop language.

This investigation aims at producing a large-coverage profile of child verb usage that can inform computational models of language acquisition, by both reporting on preferences in child language as a whole and on a developmental level. This paper is structured as follows: in section 2 we report on the resources used for this investigation, and the results are discussed in section 3. We finish with some conclusions and future work.

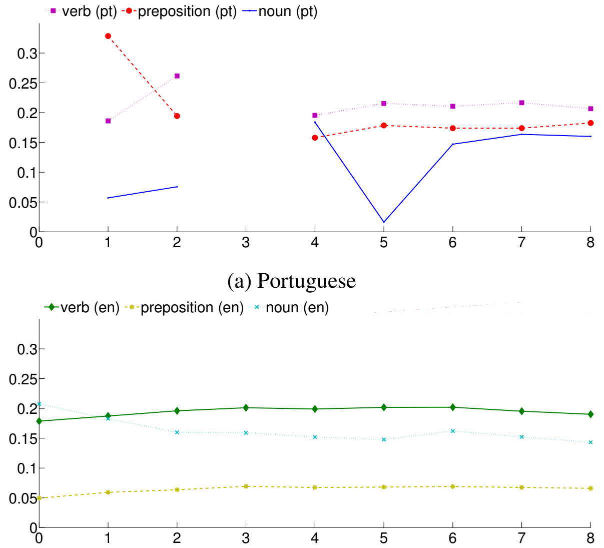
2 Resources

For examining child-produced data we use the English and Portuguese corpora from CHILDES (MacWhinney, 2000). The English corpora in CHILDES have been parsed using at least three different pipelines: MOR, POST and MEGRASP (available as part of the CHILDES distribution, the corpora are POS tagged using the MOR and POST programs (Parisse and Normand, 2000)). In addition we use a version annotated with the RASP system (Briscoe et al., 2006), that tokenizes, tags, lemmatizes and parses the input sentences, outputting syntactic trees and then adding grammatical relations (GR) as described by (Buttery and Korhonen, 2005). This corpus contains 16,649 types and 76,386,369 tokens in 3,031,217 sentences distributed by age as shown in Table 1.

Table 1: Frequency of words and sentences by age in years in CHILDES for English and Portuguese

Age	English		Portuguese	
	Words (k)	Sent (k)	Words (k)	Sent (k)
0	4,944	130	0	0
1	12,124	604	7	2
2	19,481	1,367	8	1
3	17,962	468	0	0
4	16,725	249	1	61
5	3,266	121	38	1
6	782	19	47	1
7	1,088	63	56	1
8	12	5	56	1

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.



(a) Portuguese

(b) English

Figure 1: Verbs in relation to other frequent Parts-of-speech in English (1b) and Portuguese (1a)

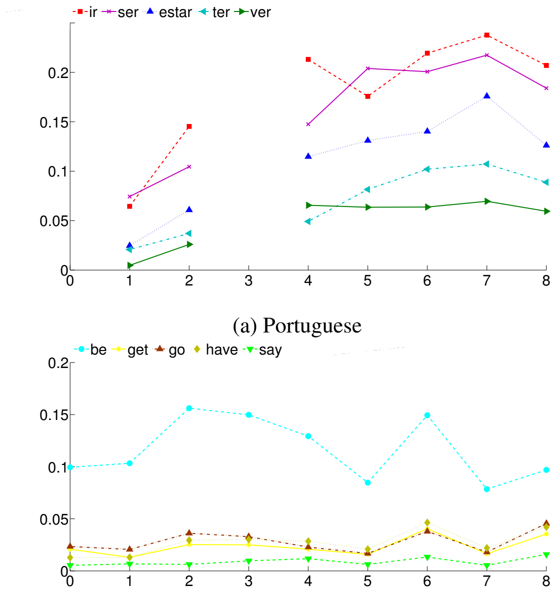
The combined size of the Portuguese corpora in sentences and words is in Table 1. These were annotated with the PALAVRAS parser, a robust parser, which has a reported accuracy of 99% for part-of-speech tagging, 96-97% for syntactic trees, and 91.8% for multiword expressions (Bick, 2000)[1]. The childes annotation were first normalized to deal with incomplete words and remove transcription annotations, and then automatically lemmatized, POS tagged, parsed and assigned semantic tags for nouns, verbs and adjectives.

3 Verbs in children data

To characterize verb usage in each of these languages we examined the distribution of verbs across the ages in terms of their relative frequencies, the number of syntactic complements with which they occur, and looking at possible links between these and age of acquisition, as reported by Gilhooly and Logie (1980).

Figure 1 focuses on the relative distributions of verbs in relation to other frequent parts-of-speech: prepositions and nouns. For both languages verbs account for around 20% of the words used, and this proportion remains constant as age increases, with the exception of the discontinuity for years 3

[1] The PALAVRAS parser was evaluated using European and Brazilian Portuguese newspaper corpora (CETENFolha and CETEMPblico) composed of 9,368 sentences.



(a) Portuguese

(b) English

Figure 2: 5 most frequent verbs in Portuguese (2a) and in English (2b)

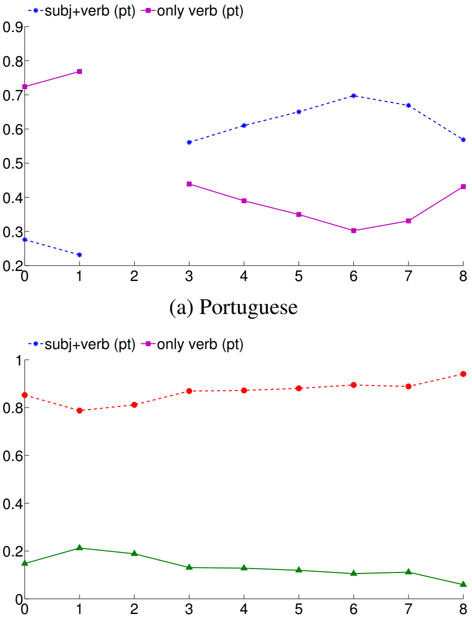
to 5 due to the lack of data for children with these ages in the Portuguese corpora in CHILDES.

Table 2: Verb types and tokens for English and Portuguese

Language	Types	Tokens
English	34,693	17,830,777
Portuguese	62,048	888,234

Table 2 shows the number of verb types and tokens in these two languages. Among these verbs, the top 5 most frequent verbs[2] for each language are: *be, get, go, have* and *say* for English and *ir* (go) *ser* (be) *estar* (be), *ter* (have) and *ver* (see) for Portuguese. These correspond to very general and polysemous verbs, and their relative proportions in the two languages remain high throughout the ages for children, figure 2. The frequencies for English are consistent with those reported by Goldberg (1999) and the Portuguese data is compatible with the crosslinguistic trends for related languages.

In terms of the syntactic characteristics of verbs in child-produced data, we examine separately



(a) Portuguese

(b) English

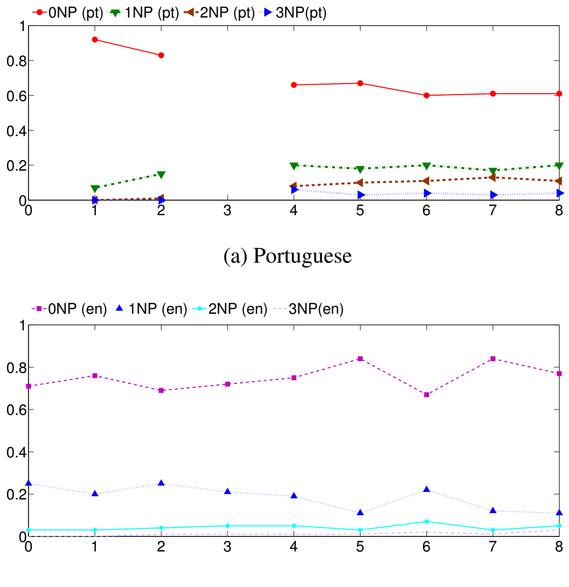
Figure 3: Percentage of sentences of verb with and without subject in Portuguese (3a) and in English (3b)

the occurrence of subjects and other complements in these languages, using the syntactic annotation provided by the RASP and PALAVRAS parsers. In the RASP annotation (Briscoe, 2006) we search for 3 types of complements in English: a direct object (dobj), the second NP complement in a double object construction (obj2) and an indirect PP object (iobj). For Portuguese, we search the PALAVRAS annotation for the following types of objects: a direct (accusative) object (ACC), a dative object (DAT), an indirect prepositional object (PIV) and an object complement (OC).[3]

For subjects figure 3 shows the occurrences of overt (subj verb) and omitted subjects (only verb) in sentences in relation to the total number of verbs (verb) for the two languages. These are a source of divergence between them as in the English data most of the verb usages consistently have an overt subject, and only around 10-20% omit the subject, but these tend to occur less as the age increases, with a peak for 2 year old children. In Portuguese, on the other hand, initially most of the verb usages omit the subject, and only later

[2] The reported frequency for each verb is for the lemmatized form, including all its inflected forms.

[3] http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html



(a) Portuguese

(b) English

Figure 4: Percentage of occurrence of objects in Portuguese (4a) and in English (4b)

this trend is reversed, but still maintaining a high proportion of subject dropping, around 40% of verb usages, and around 60% including an overt subject. The precise age for this change cannot be assessed from this data, due to the lack of sentences for 3-5 year old children in the Portuguese data. This difference between the two languages can be explained as a result of Portuguese being a (subject) pro-drop language and children being consistently exposed to subject dropping in their linguistic environment. Although English is not a pro-drop language, children, especially at an early age, still produce sentences without overt subjects, as much discussed in the literature (Valian, 1991) and more recently (Yang, 2010). Children learning pro-drop languages seem to adopt it from an early age and use it with a frequency much closer to adult usage (Valian, 1991).

In relation to other verb complements, we examine the changes in the distribution of verbs and their subcategorization frames in the corpus across children's ages. Figure 4 shows the distribution per age for verbs with one, two and three complements for both languages. As expected in general verbs with fewer complements are more frequently used and as the number of complements increases, the frequency decreases, for all ages and for both languages. Moreover, as age increases, there is a slight but constant increase in the presence of verbs with 2 and 3 complements in the corpus, with a small decrease in those with only 1, which nonetheless still account for the majority of the cases. These patterns are more clearly visible for English, as more data is available than for Portuguese for all ages.

To further investigate this we analyzed whether a relation between the number of complements of a verb and its age of acquisition could be found. For English we used the age of acquisition (AoA) scores from Gilhooly and Logie (1980) which is available for 22 of the verbs in the English data, but from these two verbs were removed from the set, as they did not occur in all the ages. For Portuguese, the scores from Marques et al. (2007) are available for only four verbs in the CHILDES corpora, and were therefore not considered in this analysis. Using the total frequency for a verb in the corpus, we calculated the relative frequencies for each number of complements (0, 1, 2 and 3) per age. For each verb and each age the number of complements with maximum frequency was used as the basis for checking if a correlation with the AoA scores for the verb could be found. In terms of the number of complements per age these verbs can be divided into 3 groups, apart from 2 of the verbs (*lock and burn*) that do not have any clear pattern:

0-obj: for verbs that are used predominantly without complements throughout the ages, *think, speak, swim, lie, turn, fly, try*;

1-obj: for verbs that appear consistently with 1 complement for all ages, *drive, chop, hate, find, win, tear*;

0-to-1: for verbs initially used mostly without complements but then consistently with 1 complement, *hurt, guess, throw, kick, hide*.

In terms of the age of acquisition, verbs in the 0-obj group tend to have lower scores than those in the second group, with a 0.72 Spearman's rank correlation coefficient indicating a high correlation between AoA and predominant number of complements of a verb. As the third group had both patterns, it was not considered in the analysis. These results suggest that the number of syntactic objects tends to increase with the age of acquisition. This may be partly explained by a potential increase in complexity as the number of obligatory arguments for a verb increase (Boynton-Hauerwas, 1998). However, more investigation is needed to confirm this trend.

4 Conclusions

In this paper we presented a wide-coverage profile of verbs in child-produced data, for English and Portuguese. We examined the distribution of some lexical and syntactic characteristics of verbs in these languages. Common trends, such as the prominent role of very general and polysemic verbs among the most frequently used and a preference for smaller number of complements were found throughout the ages in both languages. Divergences between them such as the proportion of subject dropping in each language were also found: a lower proportion for English which decreases with age and a higher proportion for Portuguese which remains relatively high. These results are compatible with those reported by e.g. Goldberg (1999) and Valian (1991), respectively. Furthermore, for English we found a high correlation between a lower age of acquisition of a verb and a lower predominant number of complements. Given the size of the Portuguese data, for some of these analyses further investigation is needed with more data to confirm the trends found.

For future work we intend to extend these analyses for other parts-of-speech, particularly nouns, also looking at other semantic and pragmatic factors, such as polysemy, concreteness and familiarity. In addition, we plan to examine intrinsic (e.g. length of words; imageability; and familiarity) and and extrinsic factors (e.g. frequency), and their effect in groups with typical development and with specific linguistic impairments.

References

Bick, E. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* [S.l.]: University of Arhus.

Bick, E. 2003. *Multi-level NER for Portuguese in a CG framework.* Proceedings of the Computational Processing of the Portuguese Language.

Boynton-Hauerwas, L. S. 1998. *The role of general all purpose verbs in language acquisition: A comparison of children with specific language impairments and their language-matched peers.* Northwestern University

Briscoe, E., Carroll, J., and Watson, R. 2006. *The second release of the rasp system.* COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Briscoe, T. 2006. *An introduction to tag sequence grammars and the RASP system parser.* Technical report in University of Cambridge, Computer Laboratory.

Buttery, P., Korhonen, A. 2005. *Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech.* Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.

Buttery, P., Korhonen, A. 2007. *I will shoot your shopping down and you can shoot all my tins– Automatic Lexical Acquisition from the CHILDES Database.* Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition. Association for Computational Linguistics.

Gilhooly, K.J. and Logie, R.H. 1980. *Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words.* Behaviour Research Methods and Instrumentation.

Goldberg, Adele E. . *The Emergence of Language*, chapter Emergence of the semantics of argument structure constructions, pages 197–212. Carnegie Mellon Symposia on Cognition Series.

Hsu, A. S., Chater, N. 2010. *Aspects of the Theory of Syntax.* MIT Press.

MacWhinney, B. 2000. *The CHILDES project: tools for analyzing talk.* Lawrence Erlbaum Associates, second edition.

Marques, J. F., Fonseca, F. L., Morais, A. S., Pinto, I. A. 2007. *Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables.* Behavior Research Methods.

Parisse, C. and Normand, M. T. Le. 2000. *Automatic disambiguation of the morphosyntax in spoken language corpora.* Behavior Research Methods, Instruments, and Computers.

Pavio, A., Yuille, J.C., and Madigan, S.A. 1968. *Concreteness, imagery and meaningfulness values for 925 words.* Journal of Experimental Psychology Monograph Supplement.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B. and Wintner, S. 2010. *Morphosyntactic annotation of CHILDES transcripts.* Journal of Child Language.

Toglia, M.P. and Battig, W.R. 1978. *Handbook of Semantic Word Norms.* New York: Erlbaum.

Valian, V. 1991. *Syntactic subjects in the early speech of American and Italian Children.* Journal of Cognition.

Yang, Charles 2010. *Three factors in language variation.* Lingua.