A large scale annotated child language construction database

Aline Villavicencio♣♠, Beracah Yankama♠, Marco A. P. Idiart♣, Robert Berwick♠

♣Federal University of Rio Grande do Sul (Brazil)
♠MIT, USA
alinev@gmail.com, beracah@mit.edu, marco.idiart@gmail.com, berwick@csail.mit.edu

Abstract

Large scale annotated corpora of child language can be of great value in assessing theoretical proposals regarding language acquisition models. For example, they can help determine whether the type and amount of data required by a proposed language acquisition model can actually be found in a naturalistic data sample. To this end, several recent efforts have augmented the CHILDES child language corpora with POS tagging and parsing information for languages such as English. With the increasing availability of robust NLP systems and electronic resources, these corpora can be further annotated with more detailed information about the properties of words, verb argument structure, and sentences. This paper describes such an initiative for combining information from various sources to extend the annotation of the English CHILDES corpora with linguistic, psycholinguistic and distributional information, along with an example illustrating an application of this approach to the extraction of verb alternation information. The end result, the English CHILDES Verb Construction Database, is an integrated resource containing information such as grammatical relations, verb semantic classes, and age of acquisition, enabling more targeted complex searches involving different levels of annotation that can facilitate a more detailed analysis of the linguistic input available to children.

Keywords: corpora, child language data, syntactic and semantic annotation

1. Introduction

Given the apparent ease with which every child acquires the language of their caretakers, human language acquisition has long been a focus of research and debate in modern cognitive science concerning the interplay of external experience and prior knowledge available to children. How much external linguistic information is actually required by children to acquire language successfully? Yet other acquisition questions refer to the developmental stages observed during acquisition and whether these can be found cross-linguistically or are language dependent. Although researchers from different areas have looked at these questions from a variety of perspectives, one common feature is the key role naturalistic data can play in order to evaluate theories and empirical predictions. Here, large-scale acquisition data can help to compare alternative theories and shed light on the practical feasibility of the conditions that they impose in order that a learner can successfully acquire language.

Consequently, collections of child-produced and child-directed speech (CDS) have been created for many languages, containing transcribed speech and in some cases even linked multimedia (audio and/or video) data. One widely used resource is the CHILDES database (MacWhinney, 1995) with transcriptions of interactions involving children of different age and language groups and from different social classes. Currently, CHILDES contains data for over 25 languages, sometimes differing as to recording periods, whether they are latitudinal or longitudinal studies, or whether a specific psycholinguistic task was involved in their collection. CHILDES is currently available in raw, part-of-speech- tagged, lemmatized and parsed formats for English (Sagae et al., 2010; Buttery and Korhonen, 2005; Buttery and Korhonen, 2007). Similar efforts have also been made for other languages, like Spanish and Hebrew (Sagae et al., 2010). Some of the databases in CHILDES also contain audio or video recordings of the interaction sessions, but these recordings for the most part remain unannotated.

The availability of resources like CHILDES have enabled large-scale investigations of both child-produced and of child-directed sentences, examining, for instance, both syntactic (Buttery and Korhonen, 2005; Buttery and Korhonen, 2007; Perfors et al., 2010; Yang, 2010; Pearl and Sprouse, 2012) and distributional (Hsu and Chater, 2010) data characteristics. Large annotated corpora like these reveal patterns in the data such as the different relative preferences in e.g. subcategorization frames for verb types in children (compared to adults) which seem to influence the way children acquire subcategorization knowledge (Buttery and Korhonen, 2007).

With the increasing availability of electronic resources and robust NLP tools, these corpora can be further annotated to add further linguistic and psycholinguistic information. This paper describes the English CHILDES Verb Construction Database (ECVCD), an initiative for combining information from various sources to extend the annotation of the CHILDES corpora, focusing on English. This involves adding to the original lexical and syntactic annotation of CHILDES information about grammatical relations, verb semantic classes, and other psycholinguistic and distributional information. The result is an integrated resource that allows complex searches involving different levels of annotation. Further, the database can be straightforwardly extended with additional annotation levels. To illustrate its potential application in language acquisition studies, we analyze the characteristics of a sample set of verbs extracted using lexical-syntactic patterns representative of certain alternation classes, which can then be used to test predictions about models of over- and under-generalization about such verb classes. In what follows, we discuss related work in section 2., the tools and resources used for the annotation in section 3., and the application of the ECVCD in section 4.. We conclude with a discussion of the implications of this initial work along with directions for future research.

2. Related Work

Language acquisition involves a complex interplay among the linguistic, distributional, and psycholinguistic characteristics of both the words and structures typically found in child-directed sentences, among many other factors. Considering words alone, some properties that seem to affect language use and recognition include intrinsic factors such as the length of a word in terms of syllables; age of acquisition; imageability; and familiarity. Additionally, extrinsic factors such as word frequency also play a role. For instance, the frequency and age of acquisition of a word seem to affect the speed of access in language and memory processes, where words that are frequent and acquired earlier in life tend to be processed faster and more accurately than those acquired later or of low frequency (Carroll and White, 1973; Morrison and Ellis, 2000). Highly frequent lexical items also seem to be the last to deteriorate in disorders such as Alzheimer's Disease (Sartori et al., 2005). This effect is consistent across languages and age groups. The choice of syntactic structures has also been linked to the properties of animacy, type, length, givenness (Marneffe et al., 2011), determinacy, and role of the constituents. These factors interact in complex ways and with different degrees of influence over the language performance of a particular individual (Marneffe et al., 2011).

Large scale annotated resources can assist in the quantitative and qualitative investigations of the linguistic, pragmatic and distributional factors that influence language acquisition and use. To consider a single current example, recently, Pearl and Sprouse (2012) used the Charniak parser along with hand-curation to annotate a portion of CHILDES to evaluate a statistical model for the acquisition of constraints on *wh*-phrase constructions. Such resources can also be used to inform the development of models for the investigation of the impact of factors for processes related to learning, aging, and cognitive impairment. In the next section we describe the development of one such dataset.

3. Tools and Resources for Annotation

The English corpora in CHILDES have been parsed using at least three different pipelines: (1) MOR, POST and MEGRASP; (2) RASP; and (3) the CHILDES Treebank. In the first, made available as part of the CHILDES distribution[1], the corpora are POS tagged using the MOR and POST programs (Parisse and Le Normand, 2000) recorded in the %mor lines, and parsed using MEGRASP (Sagae et al., 2010) with dependency parsing and grammatical relations (GRs) recorded in the %gra lines:[2]

*MOT: why don't you read ?

%mor: adv:wh|why aux|do neg|not pro|you v|read ?

%gra: 1|5|JCT 2|5|AUX 3|2|NEG 4|5|SUBJ 5|0|ROOT 6|5|PUNCT

The second pipeline is carried out using the RASP parser (Briscoe et al., 2006), which does tokenisation, tagging, lemmatization and parsing of the input sentences, outputting syntactic trees (ST) and then adding grammatical relations (GR) as described by Buttery and Korhonen (2005).[3] Each GR denotes a relation, along with its head and dependent:

*MOT: I thought we would mail it to her.

%ST: (T (S I:1 (VP think+ed:2 (S we:3 (VP would:4 mail:5 it:6 (PP to:7 she+:8))))))

%GR: (|ncsubj| |think+ed:2_VVD| |I:1_PPIS1| _)(|ccomp| _ |think+ed:2_VVD| |mail:5_VV0|)(|ncsubj| |mail:5_VV0| |we:3_PPIS2| _)(|aux| |mail:5_VV0| |would:4_VM|)(|iobj| |mail:5_VV0| |to:7_II|)(|dobj| |mail:5_VV0| |it:6_PPH1|)(|dobj| |to:7_II| |she+:8_PPHO1|)

The third approach uses the Charniak parser with Penn Treebank style part of speech tags and output, followed by hand-curation. This annotation, released in September, has not yet been evaluated for inclusion in the combined system we have constructed. Representative output from the CHILDES Treebank corpora looks like this:

(S1 (SBARQ (WHNP (WP who)) (SQ (VP (COP is) (NP (NN that)))) (. ?)))

By using annotations provided by multiple parsers, one can capitalize on the complementary strengths of each in terms of coverage and accuracy, similar to inter-annotator agreement approaches. For instance, combining these two sources of information in the search patterns one can try to maximize the precision of the results by requiring agreement between them. On the other hand, to maximize recall, the search patterns would require at least one of the sources to meet certain criteria. Moreover, as these sources may differ in terms of the precision they have for each construction, the search patterns can be optimized for prioritizing the source which produces the best accuracy for a particular case.[4]

Apart from syntactic and GR information the annotation of the verbs in each of the sentences is augmented with information about syntactic and semantic properties as defined

[1] http://childes.psy.cmu.edu/

[2] In an evaluation MEGRASP produced correct dependency relations for 96% of the relations in the gold standard, with the dependency relations being labelled with the correct GR 94% of the time.

[3] The data was kindly provided by P. Buttery and A. Korhonen and generated as described in (Buttery and Korhonen, 2005).

[4] For the target dative constructions, we manually defined the patterns using the appropriate tags for each parsers, but for a more automated process, the equivalences between the respective tagsets should also be encoded.

via Levin (1993) classes. These are formed by 190 fine-grained subclasses that classify 3,100 verb types (4,167 verb tokens) according to shared patterns of meaning and syntactic behavior. For example, the class "verbs of removing" include *delete, discharge and eject*, which express the removal of an entity from a location; their PP argument is headed by the preposition *from*. In this classification each verb can belong to more than one class, and they are all part of the annotation of a verb, as in the case of the verb *run*, that belongs to classes 26.3 (Verbs of Preparing), 47.5.1 (Swarm Verbs), 47.7 (Meander Verbs) and 51.3.2 (Run Verbs). The annotation of each verb with the classes it belongs to enables the search for sentences that belong to a given class, regardless of the verb used, and can give an indication of the distribution of verb classes according to age.

Further annotation is obtained from the MRC Psycholinguistic Database (Coltheart, 1981) which contains psychological and distributional information about words. The MRC database contains 150,837 entries with information about 26 properties, although not all properties are available for every word (e.g. IMAG is only available for 9,240 words).The following are examples of the properties it contains,where the first 3 were obtained by merging data from the norms defined by Pavio et al. (1968), Toglia and Battig (1978), and Gilhooly and Logie (1980):

- FAM - the familiarity score ranges from 100 to 700.
- CONC - the concreteness of a word ranges from 100 to 700.
- IMAG - the imageability of a word, i.e. the ease with which it allows a mental image, on a scale of 100 to 700.
- AOA - is the age of acquisition of a word from the norms of Gilhooly and Logie (1980), ranging from 100 to 700.
- NSYL - indicates the number of syllables of a word.

Some statistics about the resulting database are displayed in table 1.

Information	Sentences
Total Raw	4.84 million
MEGRASP & RASP Raw	2.5 million
MEGRASP Parsed	109,629
RASP Parsed	2.21 million
MEGRASP & RASP Parsed	98,456

Table 1: Sentences in English Corpora (UK & USA)

3.1. Annotation Schema

The annotated sentences were organized in a database, containing for each sentence the information shown in table 2. Given the focus on verbs, for search efficiency each sentence is indexed according to the verbs it contains. In addition, some words, including verbs and nouns, are further annotated with information shown in table 3 whenever it is available in the existing resources.

Fields
Sentence ID
Corpus
Speaker
File
Raw sentence
MOR and POST tags
MEGRASP dep. and GRs
RASP syntactic tree
RASP dep. and GRs
Comments

Table 2: Annotation of sentences

Fields
Word ID
Sentence ID
Levin's classes
Age of acquisition
Familiarity
Concreteness
Imageability
Number of syllables

Table 3: Annotation of Words

These levels of annotation allow for complex searches involving for example, a combination of information about a verb's lemma, target grammatical relations, and occurrence of Levin's classes in the corpora. Not all sentences have been successfully analyzed, and the comments field contains information about the missing annotations: (1) no MEGRASP parse; (2) no RASP parse. In addition, it also records cases of near perfect matches that arise from the parsers using different heuristics for e.g. non-words, meta-characters and punctuation. These required more complex matching procedures for identifying the corresponding cases in the annotations of the two parsers.

4. Verbs in Child Language

To exemplify the potential of the ECVCD for language acquisition studies, we extracted a sample set of verbs and their occurrences with double objects and prepositional dative complements and their total counts, along with information for each verb regarding membership in Levin's classes, familiarity and imageability.[5] Some of the extraction patterns used for these verb are specified in table 5. For instance, for the prepositional dative sentences, a pattern like 1 that looks for a verb that has both a direct and an indirect object headed by *to* in the RASP annotation prioritizes precision, and for extending the coverage to sentences that contain a wh-object question, pattern 2 is used. For the double object dative construction, we require strict agreement between the annotations of both RASP and MEGRASP

[5] As some of these verbs are polysemous, they are found in more than one of Levin's classes, but for this example we manually chose one of the related meanings.

which should include two objects for the target verbs as part of their GRs.

The results indicate that verbs like *read*, *speak* and *tell*, while being closely related semantically, have significantly different frequencies and subcategorization frame preferences: *read* is more frequent with a prepositional dative frame than a double object, but for *tell* the reverse is true, and *speak* is found predominantly in other subcategorization frames. In terms of familiarity, some of the most frequent verbs like *tell*, *give* and *make* all have high scores but the lowest imageability among the set of verbs. The precise role of these features and possible interactions among them for language acquisition would need to be further investigated (but see Stadthagen-Gonzalez and Davis (2006) for instance for a discussion of age of acquisition, familiarity and imageability).

5. Conclusions and future work

The development of large scale corpora of child language annotated with morphological syntactic, semantic and psycholinguistic data is of great value to both theoretical and computational research on language acquisition. Recent advances on NLP technology and an increase in the availability of linguistic and psycholinguistic resources enable the automatic addition of annotation to corpora. This paper describes the construction of the English CHILDES Verb Construction Database. It combines information from two parsing systems to capitalize on their complementary recall and precision strengths and ensure the accuracy of the searches. It also includes information about Levin's classes for verbs, and some psycholinguistic information for some of the words, like age of acquisition, familiarity and imageability, from the MRC Psycholinguistic Database. The result is a large-scale integrated resource that allows complex searches involving different annotation levels. This database can be used to inform analysis, for instance, about the complexity of the language employed with and by a child as her age increases, that can shed some light on discussions about the poverty of the stimulus. To give an indication of the search potential of the database we looked at the syntactic, semantic and distributional features of a small set of verbs concentrating on double object and prepositional dative frames.

This is an ongoing project to make the annotated data available to the research community in a user-friendly interface that allows complex patterns to be specified in a simple way. The development of the interface and evaluation with users is planned for future work. We also plan to extend the annotation adding information from other resources such as Wordnet (Fellbaum, 1998); the CHILDES Treebank; and the VALEX lexicon (Korhonen et al., 2006).

Acknowledgements

This research was partly supported by CNPq Projects 551964/2011-1, 305256/2008-4 and 309569/2009-5 and 478222/2011-4.

6. References

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

P. Buttery and A. Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

P. Buttery and A. Korhonen. 2007. I will shoot your shopping down and you can shoot all my tins: automatic lexical acquisition from the childes database. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 33–40. Association for Computational Linguistics.

J.B. Carroll and M.N. White. 1973. Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 25:85–95.

M. Coltheart. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

C. Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

K.J. Gilhooly and R.H. Logie. 1980. Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, 12:395–427.

A. S. Hsu and N. Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, 34(6):972–1016.

A. Korhonen, Y. Krymolowski, , and E. J. Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th LREC*, Genova, Italy.

B. Levin. 1993. *English verb classes and alternations - a preliminary investigation*. The University of Chicago Press.

B. MacWhinney. 1995. *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

M.C. Marneffe, S. Grimm, I. Arnon, S. Kirby, and J. Bresnan. 2011. A statistical model of grammatical choices in child production of datives sentences. *To appear in Language and Cognitive Processes*.

C.M. Morrison and A.W. Ellis. 2000. Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, 91(2):167–180.

C. Parisse and M. T. Le Normand. 2000. Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers*, 32:468–481.

A. Pavio, J.C. Yuille, and S.A. Madigan. 1968. Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement*, 76.

L. Pearl and J. Sprouse. 2012. Computational models of acquisition for islands. In J. Sprouse and N. Hornstein, editors, *Experimental Syntax and Island Effects*. Cambridge University Press.

A. Perfors, J.B. Tenenbaum, and E. Wonnacott. 2010.

Verb	Class	DObj	PObj	Total	FAM	IMG
give	Give	844	1126	12445	595	383
make	Build	154	17	20877	618	322
read	Transf. msg	69	269	5008	568	499
speak	Talk	1	0	420	600	488
tell	Transf. msg	784	32	12547	596	350

Table 4: Frames, frequency, and familiarity

<table>
<tr><th>Target</th><th>ID</th><th>Pattern</th><th>Source</th><th>Example</th></tr>
<tr><td>Prep. Dat.</td><td>1</td><td>(|iobj| |VERB| |to|) ... (|dobj| |VERB| ...)</td><td>RASP</td><td>you gave it to Daddy last night</td></tr>
<tr><td></td><td>2</td><td>(|obj| |VERB1| |What|) ...
(|xcomp| |to| |VERB1| |VERB|) (|iobj| |VERB| |to|)</td><td>RASP</td><td>What would you like to say to your father?</td></tr>
<tr><td>D. Obj. Dat.</td><td>3</td><td>(|obj2| |VERB|) (|dobj| |VERB|...)</td><td>RASP</td><td>I gave her some smarties as well, uhhuh</td></tr>
<tr><td></td><td>4</td><td>VERB|OBJ ... VERB|OBJ2</td><td>MEGRASP</td><td>I asked you a question</td></tr>
</table>

Table 5: Example search patterns

Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, (37):607–642.

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(03):705–729.

G. Sartori, L. Lombardi, and L. Mattiuzzi. 2005. Semantic relevance best predicts normal and abnormal name retrieval. *Neuropsychologia*, 43:754–770.

H. Stadthagen-Gonzalez and C. J. Davis. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38:598–605.

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

C. Yang. 2010. Three factors in language variation. *Lingua*, 120:1160–1177.