I will shoot your shopping down and you can shoot all my tins
Automatic Lexical Acquisition from the CHILDES Database

Paula Buttery and Anna Korhonen
RCEAL, University of Cambridge
9 West Road, Cambridge, CB3 9DB, UK
pjb48, alk23@cam.ac.uk

Abstract

Empirical data regarding the syntactic complexity of children's speech is important for theories of language acquisition. Currently much of this data is absent in the annotated versions of the CHILDES database. In this perliminary study, we show that a state-of-the-art subcategorization acquisition system of Preiss et al. (2007) can be used to extract large-scale subcategorization (frequency) information from the (i) child and (ii) child-directed speech within the CHILDES database without any domain-specific tuning. We demonstrate that the acquired information is sufficiently accurate to confirm and extend previously reported research findings. We also report qualitative results which can be used to further improve parsing and lexical acquisition technology for child language data in the future.

1 Introduction

Large empirical data containing children's speech are the key to developing and evaluating different theories of child language acquisition (CLA). Particularly important are data related to syntactic complexity of child language since considerable evidence suggests that syntactic information plays a central role during language acquisition, e.g. (Lenneberg, 1967; Naigles, 1990; Fisher et al., 1994).

The standard corpus in the study of CLA is the CHILDES database (MacWhinney, 2000)[1] which provides 300MB of transcript data of interactions between children and parents over 25 human languages. CHILDES is currently available in raw, part-of-speech-tagged and lemmatized formats. However, adequate investigation of syntactic complexity requires deeper annotations related to e.g. syntactic parses, subcategorization frames (SCFs), lexical classes and predicate-argument structures.

Although manual syntactic annotation is possible, it is extremely costly. The alternative is to use natural language processing (NLP) techniques for annotation. Automatic techniques are now viable, cost effective and, although not completely error-free, are sufficiently accurate to yield annotations useful for linguistic purposes. They also gather important qualitative and quantitative information, which is difficult for humans to obtain, as a side-effect of the acquisition process.

For instance, state-of-the-art statistical parsers, e.g. (Charniak, 2000; Briscoe et al., 2006), have wide coverage and yield grammatical representations capable of supporting various applications (e.g. summarization, information extraction). In addition, lexical information (e.g. subcategorization, lexical classes) can now be acquired automatically from parsed data (McCarthy and Carroll, 2003; Schulte im Walde, 2006; Preiss et al., 2007). This information complements the basic grammatical analysis and provides access to the underlying predicate-argument structure.

Containing considerable ellipsis and error, spoken child language can be challenging for current NLP techniques which are typically optimized for written adult language. Yet Sagae et al. (2005) have recently demonstrated that existing statistical parsing techniques can be usefully modified to analyse CHILDES

[1] See http://childes.psy.cmu.edu for details.

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 33–40,
Prague, Czech Republic, June 2007

with promising accuracy. Although further improvements are still required for optimal accuracy, this research has opened up the exciting possibility of automatic grammatical annotation of the entire CHILDES database in the future.

However, no work has yet been conducted on automatic acquisition of lexical information from child speech. The only automatic lexical acquisition study involving CHILDES that we are aware of is that of Buttery and Korhonen (2005). The study involved extracting subcategorization information from (some of) the adult (child-directed) speech in the database, and showing that this information differs from that extracted from the spoken part of the British National Corpus (BNC) (Burnard, 1995).

In this paper, we investigate whether state-of-the-art subcategorization acquisition technology can be used—without any domain-specific tuning—to obtain large-scale verb subcategorization frequency information from CHILDES which is accurate enough to show differences and similarities between child and adult speech, and thus be able to provide support for syntactic complexity studies in CLA.

We use the new system of Preiss et al. (2007) to extract SCF frequency data from the (i) child and (ii) child-directed speech within CHILDES. We show that the acquired information is sufficiently accurate to confirm and extend previously reported SCF (dis)similarities between the two types of data. In particular, we demonstrate that children and adults have different preferences for certain types of verbs, and that these preferences seem to influence the way children acquire subcategorization. In addition, we report qualitative results which can be used to further improve parsing and lexical acquisition technology for spoken child language data in the future.

2 Subcategorization Acquisition System

We used for subcategorization acquisition the new system of Preiss, Briscoe and Korhonen (2007) which is essentially a much improved and extended version of Briscoe and Carroll's (1997) system. It incorporates 168 SCF distinctions, a superset of those found in the COMLEX Syntax (Grishman et al., 1994) and ANLT (Boguraev et al., 1987) dictionaries. Currently, SCFs abstract over specific lexically governed particles and prepositions and specific predicate selectional preferences but include some derived semi-predictable bounded dependency constructions, such as particle and dative movement—this will be revised in future versions of the SCF system.

The system tokenizes, tags, lemmatizes and parses input sentences using the recent (second) release of the RASP (Robust Accurate Statistical Parsing) system (Briscoe et al., 2006) which parses arbitrary English text with state-of-the-art levels of accuracy. SCFs are extracted from the grammatical relations (GRs) output of the parser using a rule-based classifier. This classifier operates by exploiting the close correspondence between the dependency relationships which the GRs embody and the head-complement structure which subcategorization acquisition attempts to recover. Lexical entries of extracted SCFs are constructed for each word in the corpus data. Finally, the entries may be optionally filtered to obtain a more accurate lexicon. This is done by setting empirically determined thresholds on the relative frequencies of SCFs.

When evaluated on cross-domain corpora containing mainly adult language, this system achieves 68.9 F-measure[2] in detecting SCF types—a result which compares favourably to those reported with other comparable SCF acquisition systems.

3 Data

The English (British and American) sections of the CHILDES database (MacWhinney, 2000) were used to create two corpora: 1) CHILD and 2) CDS. Both corpora contained c. 1 million utterances which were selected from the data after some utterances containing un-transcribable sections were removed. Speakers were identified using speaker-id codes within the CHAT transcriptions of the data:[3] CHILD contained the utterances of speakers identified as target children; CDS contained input from speakers identified as parents/caretakers. The mean utterance length (measured in words) in CHILD and CDS were 3.48 and 4.61, respectively. The mean age of the child speaker in CHILD is around 3 years 6 months.[4]

[2]See Section 4 for details of F-measure.

[3]CHAT is the transcription and coding format used by all the transcriptions within CHILDES.

[4]The complete age range is from 1 year and 1 month up to 7 years.

3.1 Test Verbs and SCF Lexicons

We selected a set of 161 verbs for experimentation. The words were selected at random, subject to the constraint that a sufficient number of SCFs would be extracted (> 100) from both corpora to facilitate maximally useful comparisons. All sentences containing an occurrence of one of the test verbs were extracted from the two corpora and fed into the SCF acquisition system described earlier in section 2.

In some of our experiments the two lexicons were compared against the VALEX lexicon (Korhonen et al., 2006)—a large subcategorization lexicon for English which was acquired automatically from several cross-domain corpora (containing both written and spoken language). VALEX includes SCF and frequency information for 6,397 English verbs. We employed the most accurate version of the lexicon here (87.3 F-measure)—this lexicon was obtained by selecting high frequency SCFs and supplementing them with lower frequency SCFs from manually built lexicons.

4 Analysis

4.1 Methods for Analysis

The similarity between verb and SCF distributions in the lexicons was examined. To maintain a robust analysis in the presence of noise, multiple similarity measures were used to compare the verb and SCF distributions (Korhonen and Krymolowski, 2002). In the following $p = (p_i)$ and $q = (q_i)$ where p_i and q_i are the probabilities associated with SCF$_i$ in distributions (lexicons) P and Q:

- Intersection (IS) - the intersection of non-zero probability SCFs in p and q;
- Spearman rank correlation (RC) - lies in the range $[1; 1]$, with values near 0 denoting a low degree of association and values near -1 and 1 denoting strong association;
- Kullback-Leibler (KL) distance - a measure of the additional information needed to describe p using q, KL is always ≥ 0 and $= 0$ only when $p \equiv q$;

The SCFs distributions acquired from the corpora for the chosen words were evaluated against: (i) a gold standard SCF lexicon created by merging the SCFs in the COMLEX and ANLT syntax dictionaries—this enabled us to determine the accuracy of the acquired SCFs; (ii) another acquired SCF lexicon (as if it were a gold standard)—this enabled us to determine similarity of SCF types between two lexicons. In each case

Verb	CHILD	CDS
go	1	1
want	2	2
get	3	3
know	4	4
put	5	6
see	6	5
come	7	10
like	8	7
make	9	11
say	10	8
take	11	13
eat	12	14
play	13	15
need	14	16
look	15	12
fall	16	22
sit	17	21
think	18	9
break	19	27
give	20	17

Table 1: Ranks of the 20 most frequent verbs in CHILD and in CDS

we recorded the number of *true positives* (TPs), correct SCFs, *false positives* (FPs), incorrect SCFs, and *false negatives* (FNs), correct SCFs not in the gold standard.

Using these counts, we calculated type precision (the percentage of SCF types in the acquired lexicon which are correct), type recall (the percentage of SCF types in the gold standard that are in the lexicon) and F-measure:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (1)$$

4.2 Verb Analysis

Before conducting the SCF comparisons we first compared (i) our 161 test verbs and (ii) all the 1212 common verbs and their frequencies in CHILD and CDS using the Spearman rank correlation (RC) and the Kullback-Leibler distance (KL). The result was a strong correlation between the 161 test verbs (RC = 0.920 $\pm$ 0.0791, KL = 0.05) as well as between all the 1212 verbs (RC = 0.851 $\pm$ 0.0287, KL = 0.07) in the two corpora.

These figures suggest that the child-directed speech (which is less diverse in general than speech between adults, see e.g. the experiments of Buttery and Korhonen (2005)) contains a very similar distribution of verbs to child speech. This is to be expected since the corpora essentially contain separate halves of the same interactions.

However, our large-scale frequency data makes it possible to investigate the cause for the apparently small differences in the distributions. We did this by examining the strength of correlation throughout the ranking. We compared the ranks of the individual verbs and discovered that the most frequent verbs in the two corpora have indeed very similar ranks. Table 1 lists the 20 most frequent verbs in CHILD (starting from the highest ranked verb) and shows their ranks in CDS. As illustrated in the table, the top 4 verbs are identical in the two corpora (*go, want, get, know*) while the top 15 are very similar (including many action verbs e.g. *put, look, sit, eat,* and *play*).

Yet some of the lower ranked verbs turned out to have large rank differences between the two corpora. Two such relatively highly ranked verbs are included in the table—*think* which has a notably higher rank in CDS than in CHILD, and *break* which has a higher rank in CHILD than in CDS. Many other similar cases were found in particular among the medium and low frequency verbs in the two corpora.

To obtain a better picture of this, we calculated for each verb its rank difference between CHILD vs. CDS. Table 2 lists 40 verbs with substantial rank differences between the two corpora. The first column shows verbs which have higher ranks in CHILD than in CDS, and the second column shows verbs with higher ranks in CDS than in CHILD. We can see e.g. that children tend to prefer verbs such as *shoot, die* and *kill* while adults prefer verbs such as *remember, send* and *learn*.

To investigate whether these differences in preferences are random or motivated in some manner, we classified the verbs with the largest differences in ranks (>10) into appropriate Levin-style lexical-semantic classes (Levin, 1993) according to their predominant senses in the two corpora.[5] We discovered that the most frequent classes among the verbs that children prefer are HIT (e.g. *bump, hit, kick*), BREAK (e.g. *crash, break, rip*), HURT (e.g. *hurt, burn, bite*) and MOTION (e.g. *fly, jump, run*) verbs. Overall, many of the preferred verbs (regardless of the class) express negative actions or feelings (e.g. *shoot, die, scare, hate*).

[5]This classification was done manually to obtain a reliable result.

CHILD		CDS	
shoot	*tie*	*remember*	*hope*
hate	*wish*	*send*	*suppose*
die	*cut*	*learn*	*bet*
write	*crash*	*wipe*	*kiss*
use	*kick*	*pay*	*smell*
bump	*scare*	*feed*	*guess*
win	*step*	*ask*	*change*
lock	*burn*	*feel*	*set*
fight	*stand*	*listen*	*stand*
jump	*care*	*wait*	*wonder*

Table 2: 20 verbs ranked higher in (i) child speech and (ii) child-directed speech.

In contrast, adults have a preference for verbs from classes expressing cognitive processes (e.g. *remember, suppose, think, wonder, guess, believe, hope, learn*) or those that can be related to the education of children, e.g. the WIPE verbs *wash, wipe* and *brush* and the PERFORMANCE verbs *draw, dance* and *sing*. In contrast to children, adults prefer verbs which express positive actions and feelings (e.g. *share, help, love, kiss*).

It is commonly reported that child CLA is motivated by a wish to communicate desires and emotions, e.g. (Pinker, 1994), but a relative preference in child speech over child-directed speech for certain verb types or verbs expressing negative actions and feelings has not been explicitly shown on such a scale before. While this issue requires further investigation, our findings already demonstrate the value of using large scale corpora in producing novel data and hypotheses for research in CLA.

4.3 SCF Analysis

4.3.1 Quantitative SCF Comparison

The average number of SCFs taken by studied verbs in the two corpora proved quite similar. In unfiltered SCF distributions, verbs in CDS took on average a larger number of SCFs (29) than those in CHILD (24), but in the lexicons filtered for accuracy the numbers were identical (8–10, depending on the filtering threshold applied). The intersection between the CHILD / CDS SCFs and those in the VALEX lexicon was around 0.5, indicating that the two lexicons included only 50% of the SCFs in the lexicon extracted from general (cross-domain) adult language corpora. Recall against VALEX was consequently low (between 48% and 68% depending on the filtering threshold) but precision was around 50-60% for both CHILDES and CDS lexicons

Measures	Unfilt.	Filt.
Precision (%)	82.9	88.7
Recall (%)	69.3	44.5
F-measure	75.5	59.2
IS	0.73	0.62
RC	0.69	0.72
KL	0.33	0.46

Table 3: Average results when SCF distributions in CHILD and CDS are compared against each other.

(also depending on the filtering threshold), which is a relatively good result for the challenging CHILDES data. However, it should be remembered that with this type of data it would not be expected for the SCF system to achieve as high precision and recall as it would on, for instance, adult written text and that the missing SCFs and/or misclassified SCFs are likely to provide us with the most interesting information.

As expected, there were differences between the SCF distributions in the two lexicons. Table 3 shows the results when the CHILD and CDS lexicons are compared against each other (i.e. using the CDS as a gold standard). The comparison was done using both the unfiltered and filtered (using relative frequency threshold of 0.004) versions of the lexicons. The similarity in SCF types is 75.5 according to F-measure in the unfiltered lexicons and 59.2 in filtered ones.[6]

4.3.2 Qualitative SCF Comparison

Our qualitative analysis of SCFs in the two corpora revealed reasons for the differences. Table 4 lists the 10 most frequent SCFs in CHILD (starting from the highest ranked SCF), along with their ranks in CDS and VALEX. The top 3 SCFs (NP, INTRANSITIVE and PP frames) are ranked quite similarly in all the corpora. Looking at the top 10 SCFs, CHILD appears, as expected, more similar to CDS than with VALEX, but large differences can be detected in lower ranked frames.

To identify those frames, we calculated for each SCF its difference in rank between CHILD vs. CDS. Table 5 exemplifies some of the SCFs with the largest rank differences. Many of these concern frames involving sentential complementation. Children use more frequently than adults SCFs involving THAT and HOW complementation, while adults have a preference for SCFs involving WHETHER, ING and IF complementation.

Although we have not yet looked at SCF differences across ages, these discoveries are in line with previous findings, e.g. (Brown, 1973), which indicate that children master the sentential complementation SCFs preferred by adults (in our experiment) fairly late in the acquisition process. With a mean utterance length for CHILD at 3.48, we would expect to see relatively few of these frames in the CHILD corpus—and consequently a preference for the simpler THAT constructions.

4.4 The Impact of Verb Type Preferences on SCF Differences

Given the new research findings reported in Section 4.2 (i.e. the discovery that children and adults have different preferences for many medium-low frequency verbs) we investigated whether verb type preferences play a role in SCF differences between the two corpora. We chose for experimentation 10 verbs from 3 groups:

1. Group 1 – verbs with similar ranks in CHILD and CDS: *bring, find, give, know, need, put, see, show, tell, want*
2. Group 2 – verbs with higher ranks in CDS: *ask, feel, guess, help, learn, like, pull, remember, start, think*
3. Group 3 – verbs with higher ranks in CHILD: *break, die, forget, hate, hit, jump, scare, shoot, burn, wish*

The test verbs were selected randomly, subject to the constraint that their absolute frequencies in the two corpora were similar.[7] We first correlated the unfiltered SCF distributions of each test verb in the two corpora against each other and calculated the similarity in the SCF types using the F-measure. We then evaluated for each group, the accuracy of SCFs in unfiltered distributions against our gold standard (see Section 4.1). Because the gold standard was too ambitious in terms of recall, we only calculated the precision figures: the average number of TP and FP SCFs taken by test verbs.

The results are included in Table 6. Verbs in Group 1 show the best SCF type correlation (84.7 F-measure) between the two corpora although they are the richest in terms of subcategorization (they take the highest number of SCFs out of the three groups). The SCF correlation is clearly lower in Groups 2 and 3, although

[6]The fact that the unfiltered lexicons appear so much more similar suggests that some of the similarity is due to similarity in incorrect SCFs (many of which are low in frequency, i.e. fall under the threshold).

[7]This requirement was necessary because frequency may influence subcategorization acquisition performance.

SCF	Example sentence	CHILD	CDS	VALEX
NP	*I love rabbits*	1	1	1
INTRANS	*I sleep with a pillow and blanket*	2	2	2
PP	*He can jump over the fence*	3	4	3
PART	*I can't give up*	4	7	9
TO-INF-SC	*I want to play with something else*	5	3	6
PART-NP/NP-PART	*He looked it up*	6	6	7
NP-NP	*Ask her all these questions*	7	5	18
NP-INF-OC	*Why don't you help her put the blocks in the can ?*	8	9	60
INTR-RECIP	*So the kitten and the dog won't fight*	9	8	48
NP-PP	*He put his breakfast in the bin*	10	10	4

Table 4: 10 most frequent SCFs in CHILD, along with their ranks in CDS and VALEX.

	SCF	Example sentence
CHILD	MP	*I win twelve hundred dollars*
	INF-AC	*You can help me wash the dishes*
	PP-HOW-S	*He explained to her how she did it*
	HOW-TO-INF	*Daddy can you tell me how to spell Christmas carols?*
	NP-S	*He did not tell me that it was gonna cost me five dollars*
CDS	ING-PP	*Stop throwing a tantrum*
	NP-AS-NP	*I sent him as a messenger*
	NP-WH-S	*I'll tell you whether you can take it off*
	IT WHS, SUBTYPE IF	*How would you like it if she pulled your hair?*
	NP-PP-PP	*He turned it from a disaster into a victory*

Table 5: Typical SCFs with higher ranks in (i) CHILD and (ii) CDS.

	Measures	Group1	Group2	Group3
SCF similarity	F-measure	84.7	72.17	75.60
SCF accuracy	TPs CDS	12	11	7
	TPs CHILD	10	9	8
	FPs CDS	36	29	13
	FPs CHILD	32	18	15

Table 6: Average results for 3 groups when (i) unfiltered SCF distributions in CHILD and CDS are compared against each other (SCF similarity) and when (ii) the SCFs in the distributions are evaluated against a gold standard (SCF accuracy).

the verbs in these groups take fewer SCFs. Interestingly, Group 3 is the only group where children produce more TPs and FPs on average than adults do, i.e. both correct and incorrect SCFs which are not exemplified in the adult speech. The frequency effects controlled, the reason for these differences is likely to lie in the differing relative preferences children and adults have for verbs in groups 2 and 3, which we think may impact the richness of their language.

4.5 Further Analysis of TP and FP Differences

We looked further at the interesting TP and FP differences in Group 3 to investigate whether they tell us something about (i) how children learn SCFs (via both TPs and FPs), and (ii) how the parsing / SCF extraction system could be improved for CHILDES data in the future (via the FPs).

We first made a quantitative analysis of the relative difference in TPs and FPs for all the SCFs in both corpora. The major finding of this high level analysis was a significantly high FP rate for some ING frames (e.g. PART-ING-SC, ING-NP-OMIT, NP-ING-OC) within CHILD (e.g. *"car going hit"*, *"I hurt hand moving"*). This agrees with many previous studies, e.g. (Brown, 1973), which have shown that children overextend and incorrectly use the "ing" morpheme during early acquisition.

A qualitative analysis of the verbs from Group 3 was then carried out, looking for the following scenarios:

- SCF is a FP in both CHILD and CDS - either i) the gold standard is incomplete, or ii) there is error in the parser/subcategorization system with respect to the CHILDES domain.

- SCF is a TP in CDS and not present in CHILD - children have not acquired the frame despite exposure to it (perhaps it is complicated to acquire).

- SCF is a TP in CHILD but not present in CDS - adults are not using the frame but the children have acquired it. This indicates that either i) children are acquiring the frame from elsewhere in their environment (perhaps from a television),

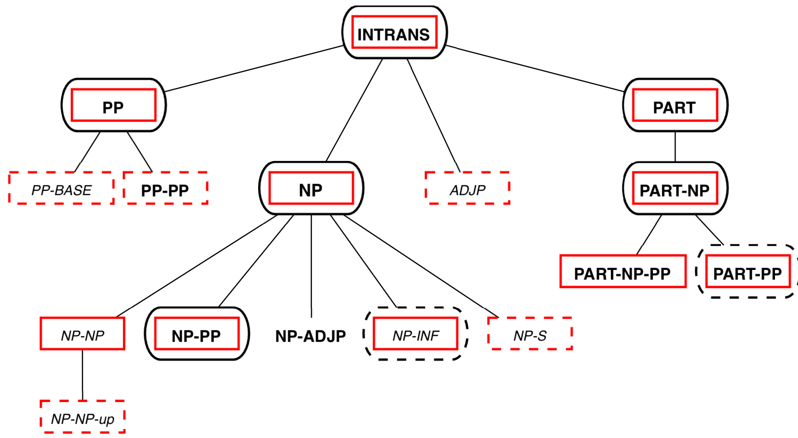

Figure 1: SCFs obtained for the verb *shoot*

or ii) there is a misuse of the verb's semantic class in child speech.

- SCF is a FP in CHILD but not present in CDS - children should not have been exposed to this frame but they have acquired it. This indicates either i) a misuse of the verb's semantic class, or ii) error in the parsing/subcategorization technology with respect to the child-speech domain.

These scenarios are illustrated in Figure 1 which graphically depicts the differences in TPs and FPs for the verb *shoot*. The SCFs have been arranged in a complexity hierarchy where complexity is defined in terms of increasing argument structure.[8] SCFs found within our ANLT-COMLEX gold standard lexicon for *shoot* are indicated in bold-face. A right-angled rectangle drawn around a SCF indicates that the frame is present in CHILD—a solid line indicating a strong presence (relative frequency $>$ 0.010) and a dotted line indicating a weak presence (relative frequency $>$ 0.005). Rounded-edge rectangles represent the presence of SCFs within CDS similarly. For example, the frame NP represents a TP in both CHILD and CDS and the frame NP-NP represents a FP within CHILD.

With reference to Figure 1, we notice that all of the SCFs present in CHILD are directly connected within the hierarchy and there is a tendency for weakly present SCFs to inherit from those strongly present. A possible explanation for this is that children are exploring SCFs—trying out frames that are slightly more complex than those already acquired (for a learning algorithm that exploits such a hypothesis in general see (Buttery, 2006)).

The SCF NP-NP is strongly present in CHILD despite being a FP. Inspection of the associated utterances reveals that some instances NP-NP are legitimate but so uncommon in adult language that they are omitted from the gold-standard (e.g. *"can i shoot us all to pieces"*. However, other instances demonstrate a misunderstanding of the semantic class of the verb; there is possible confusion with the semantic class of *send* or *throw* (e.g. *"i shoot him home"*).

The frame NP-INF is a FP in both corpora and a frequent FP in CHILD. Inspection of the associated utterances flags up a parsing problem. Frame NP-INF can be illustrated by the sentences *"he helped her bake the cake"* or *"he made her sing"*, however, within CHILD the NP-INF has been acquired from utterances such as *"i want ta shoot him"*. The RASP parser has mistagged the word *"ta"* leading to a misclassification by the SCF extraction system. This problem could be solved by augmenting RASP's current grammar with a lexical entry specifying *"ta"* as an alternative to infinitival *"to"*.

In summary, our analysis of TP and FP differences has confirmed previous studies regarding the nature of child speech (the over-extension of the *"ing"* morpheme). It has also demonstrated that TP/FP analysis can be a useful diagnostic for parsing/subcategorization extraction problems within a new data domain. Further, we suggest that analysis of FPs can provide empirical data regarding the manner in which children learn the semantic classes of

[8] For instance, the intransitive frame INTRANS is less complex than the transitive frame NP, which in turn is less complex than the di-transitive frame NP-NP. For a detailed description of all SCFs see (Korhonen, 2002).

verbs (a matter that has been much debated e.g. (Levin, 1993), (Brooks and Tomasello, 1999)).

5 Conclusion

We have reported the first experiment for automatically acquiring verbal subcategorization from both child and child-directed parts of the CHILDES database. Our results show that a state-of-the-art subcategorization acquisition system yields useful results on challenging child language data even without any domain-specific tuning. It produces data which is accurate enough to confirm and extend several previous research findings in CLA. We explore the discovery that children and adults have different relative preferences for certain verb types, and that these preferences influence the way children acquire subcategorization. Our work demonstrates the value of using NLP technology to annotate child language data, particularly where manual annotations are not readily available for research use. Our pilot study yielded useful information which will help us further improve both parsing and lexical acquisition performance on spoken/child language data. In the future, we plan to optimize the technology so that it can produce higher quality data for investigation of syntactic complexity in this domain. Using the improved technology we plan to then conduct a more thorough investigation of the interesting CLA topics discovered in this study—first concentrating on SCF differences in child speech across age ranges.

References

B. Boguraev, J. Carroll, E. J. Briscoe, D. Carter, and C. Grover. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proc. of the 25th Annual Meeting of ACL*, pages 193–200, Stanford, CA.

E Briscoe and J Carroll. 1997. Automatic extraction of subcategorization from corpora. In *5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC. ACL.

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

P Brooks and M Tomasello. 1999. Young children learn to produce passives with nonce verbs. *Developmental Psychology*, 35:29–44.

R Brown. 1973. *A first Language: the early stages*. Harvard University Press, Cambridge, MA.

L. Burnard, 1995. *The BNC Users Reference Guide*. British National Corpus Consortium, Oxford, May.

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*, Saarbrucken, Germany.

P Buttery. 2006. *Computational Models for First Language Acquisition*. Ph.D. thesis, University of Cambridge.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA.

C. Fisher, G. Hall, S. Rakowitz, and L. Gleitman. 1994. When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1–4):333–375, April.

R. Grishman, C. Macleod, and A. Meyers. 1994. COMLEX Syntax: Building a Computational Lexicon. In *Proc. of COLING*, Kyoto.

A. Korhonen and Y. Krymolowski. 2002. On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In *Proc. of the 6th CoNLL*, pages 91–97, Taipei, Taiwan.

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

A Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge. Thesis published as Technical Report UCAM-CL-TR-530.

E Lenneberg. 1967. *Biological Foundations of Language*. Wiley Press, New York, NY.

B Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press, Chicago, IL.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.

D. McCarthy and J. Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4).

L Naigles. 1990. Children use syntax to learn verb meanings. *Journal of Child Language*, 17:357–374.

S Pinker. 1994. *The Language Instinct: How the Mind Creates Language*. Harper Collins, New York, NY.

J. Preiss, E. J. Briscoe, and A. Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of ACL*, Prague, Czech Republic. To appear.

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child langugage. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan.

S. Schulte im Walde. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.