Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech

Paula Buttery and Anna Korhonen
Natural Language and Information Processing Group
Computer Laboratory, Cambridge University
15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK
{*paula.buttery, anna.korhonen*}*@cl.cam.ac.uk*

Abstract

Empirical data regarding the syntactic complexity of child directed speech (CDS) is necessary for determining its rôle in language acquisition. Of particular importance is data related to the predicate-argument structures and verb subcategorization frames (SCFs). However, manual analysis of SCFs is costly and consequently available data for evaluating theories is sparse. We address this problem by using the most comprehensive subcategorization system available to automatically acquire large scale empirical data related to verb SCFs from CDS (an edited corpus of the CHILDES database (MacWhinney, 1995)). We compare this data against adult speech (a subset of the spoken part of the British National Corpus (BNC) (Leech, 1992)) and find that SCFs typical to CDS are different and often simpler than those typical to speech between adults. We discuss the impact of our findings on the prevailing theories of language acquisition.

1 Introduction

Understanding the rôle, if any, of child directed speech (CDS) is of fundamental importance to language acquisition. Several manual small scale studies (see Snow (1986) for an overview) have suggested that CDS is very different from speech between adults: intonation is often exaggerated, a specific vocabulary can be used, and sometimes even specific syntactic structures. However, the rôle of CDS is by no means clear. Pine (1994), amongst others, speculates that the purpose of CDS is to merely engage the child in conversation. Snow (1986), on the other hand, suggests that CDS is actually teaching the child language. Clearly, larger-scale studies into the nature of CDS are required before we can begin to establish its rôle in acquisition. This paper details a systematic, large-scale investigation into the syntactic properties of verbs in CDS.

There is considerable evidence that syntactic information, in particular, is informative during language acquisition (e.g. (Lenneberg, 1967), (Naigles, 1990) and (Fisher *et al.*, 1994)). Often theories rely on syntactic diversity in the child's input for successful acquisition. For example, Landau and Gleitman (1985) suggest that children use verb subcategorization frames (SCFs) to identify novel word meanings; arguing that in many cases surface-structure/situation pairs are insufficient or even misleading about a verb's interpretation . Consider the sentences *Did you eat your cookie?* and *Do you want your cookie?* According to Landau and Gleitman the SCFs of *eat* and *want* cue their interpretations, i.e. *want* occurs with sentential complements, suggesting a mental component to its interpretation. Furthermore, they suggest that SCFs provide convergent evidence on the meaning of a verb. For instance, if *John zirks bill the book* the learner assumes *zirk* to be an active verb of transfer (such as *bring, throw, explain*), whereas if *John is zirking that the book is dull* the learner interprets *zirk* to be a mental verb.

Such a syntactically intensive theory of acquisition can only be supported if the input to children is sufficiently complex and diverse in its SCFs. In general, CDS is thought to be syntactically simpler than adult speech, using simpler and fewer SCFs (Snow, 1986). If the rôle of CDS is to teach language, as Snow suggests, then we may have a conflict with acquisition theories that require syntactic complexity and diversity.

Manual analysis of SCFs is very costly and therefore not ideal for large scale studies in specific domains, such as CDS. Automatic acquisition of SCFs from corpora now produces fairly accurate lexical data useful for (psycho)linguistic research (e.g. Roland et al. (2000)). However, these methods are yet to be applied to CDS.

In this paper, we address the problem by using the most comprehensive subcategorization system available for English to automatically acquire large scale empirical data related to verb SCFs from CDS. We use both qualitative and quantitative methods

to compare the resulting data against that obtained from a corpus of adult speech. We discuss our findings in relation to the prevailing theories of language acquisition.

Section 2 describes our method for subcategorization acquisition and section 3 introduces the corpora we used in our work. Our experiments and results are reported in section 4 and section 5 provides discussion and summarises our observations.

2 Methodology

We used for subcategorization acquisition the latest version of Briscoe and Carroll's (1997) system (Korhonen, 2002) which incorporates 163 SCF distinctions, a superset of those found in the ANLT (Boguraev *et al.*, 1987) and COMLEX (Grishman *et al.*, 1994) dictionaries. The 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.

The system first extracts sentences containing specific predicates from a corpus. The resulting data is tagged, lemmatized and parsed using the 'RASP' system (Robust Accurate Statistical Parser; (Briscoe and Carroll, 2002)). Local syntactic frames including the syntactic categories and head lemmas of constituents are then extracted from parses. The resulting patterns are classified to SCFs on the basis of the feature values of syntactic categories and the head lemmas in each pattern. Finally a lexical entry is constructed for each verb and SCF combination whose relative frequency is higher than an empirically defined threshold.

3 Corpora

In order to make valid comparisons between SCF frequencies in CDS against adult speech there is a necessity to first ensure that the corpora are controlled for all other variables. Roland and Jurafsky (1998) have shown that there are subcategorization differences between written and spoken corpora, and furthermore that subcategorization is affected by genre and discourse type. Hence, we use only spoken data for both corpora and restrict data to conversation between family members and friends.

To ensure sufficient data for subcategorization acquisition, we have had to use an American English source for the CDS corpus although we had a British English source for the adult speech corpus. However, we do not expect this to be a problem: Roland *et al* (2000) have shown that subcategorization probabilities are fairly stable across American vs. British English corpora; finding any exceptions to be the result of subtle shifts in verb sense due to genre.

The following sections describe the two corpora we chose to experiment with.

3.1 Child Directed Speech - CHILDES Corpus

The CDS corpus has been created from several sections of the CHILDES database (MacWhinney, 1995): Demetras1 (Demetras, 1989b); Demetras2 (Demetras, 1989a); Higginson (Higginson, 1985); Post (Post, 1992); Sachs (Sachs, 1983); Suppes (Suppes, 1974); Warren-Leubecker (Warren-Leubecker, 1982). These sections of the database exhibit naturalistic interactions between a child and caretaker (average child age 2;7). Speakers are both male and female, from a variety of backgrounds and from several locations around the USA. Child speech has been removed from the corpus and there is no reading. The corpus contains 534,782 words and has an average utterance length of 4.8.

3.2 Adult Speech - BNC Corpus

Our adult speech corpus has been manually constructed from the demographic part of the spoken British National Corpus (BNC) (Leech, 1992) such that it contains friend/family interactions where no children were present. The speakers were recruited by the British Market Research Bureau and come from a variety of social backgrounds. Speakers are both male and female, from several locations around the UK and all have an age of at least 15. Conversations were recorded unobtrusively over two or three days, and details of each conversation were logged. The corpus contains 835,461 words and has an average utterance length of 7.3.

4 Analysis

4.1 SCF Lexicons

We took the two corpora and extracted from them up to a maximum of 5000 utterances per verb. To make the results comparable, an equal number of utterances per verb were used for both corpora. In practice this number was often determined by CHILDES, which was smaller of the two corpora. It was also affected by the highly zipfian nature of verb distributions (see e.g. Korhonen (2002)), i.e. the fact that most verb types are extremely infrequent in language.

4.2 Methods for Analysis

Both qualitative and quantitative methods were used to compare the data in two SCF lexicons. The similarity between SCF distributions in the lexicons was

examined using various measures of distributional similarity. These include:

- Kullback-Leibler distance—a measure of the additional information needed to describe p using q, KL is always ≥ 0 and $= 0$ only when $p \equiv q$;
- Jenson-Shannon divergence—a measure which relies on the assumption that if p and q are similar, they are close to their average;
- Cross entropy—a measure of the information need to describe a true distribution p using a model distribution q, cross entropy is minimal when p and q are identical;
- Skew divergence—smooths q by mixing with p;
- Rank correlation—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;
- Intersection—the intersection of non-zero probability SCFs in p and q;

where p and q are the distributions of SCFs in lexicons P and Q. For details of these measures see Korhonen and Krymolowski (2002).

In some of our experiments, the acquired SCFs were contrasted against a gold standard SCF lexicon created by merging the SCFs in the COMLEX and ANLT syntax dictionaries. We did this by calculating type precision (the percentage of SCF types that the system proposes which are correct), type recall (the percentage of SCF types in the gold standard that the system proposes) and F-measure:

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

4.3 Difference in Verb Types

Before conducting the SCF comparisons, we examined the 100 most frequent verbs in the BNC corpus versus the CHILDES corpus to get a more complete picture of the differences between the two data. We discovered that some verbs tend to be frequent in both corpora, e.g. *go, get, think, like, make, come, take*. However, closer analysis of the data revealed large differences. We discovered that in general, action verbs (e.g. *put, look, let, sit, eat, play*) are more frequent in CHILDES, while mental state verbs (e.g. *say, know, mean, suppose, ask, feel, seem*)—which tend to have richer argument structure—are more frequent in BNC. The 30 most frequent verbs in the two corpora are listed in Figure 1, in the order of their frequency, starting from the highest ranked.

Rank	BNC	n	CHILDES	n
1	*get*	5000+	*go*	5000+
2	*go*	5000+	*be*	5000+
3	*say*	5000+	*do*	5000+
4	*be*	5000+	*see*	4200
5	*know*	5000+	*put*	4037
6	*do*	5000+	*get*	4018
7	*think*	4074	*want*	3411
8	*see*	2852	*can*	3409
9	*like*	2827	*let*	2771
10	*can*	2710	*look*	2585
11	*come*	2602	*think*	2280
12	*want*	2148	*like*	2038
13	*mean*	2078	*know*	1768
14	*look*	1930	*say*	1755
15	*put*	1776	*come*	1693
16	*take*	1443	*make*	1692
17	*tell*	1122	*okay*	1593
18	*make*	1092	*take*	1356
19	*use*	1016	*eat*	1172
20	*will*	1007	*give*	990
21	*give*	920	*play*	944
22	*buy*	590	*tell*	860
23	*leave*	548	*find*	661
24	*keep*	545	*happen*	581
25	*pay*	543	*sit*	580
26	*let*	536	*read*	571
27	*remember*	517	*remember*	563
28	*work*	495	*try*	556
29	*suppose*	489	*fall*	546
30	*play*	477	*will*	537

Figure 1: 30 most frequent verbs in adult speech (BNC) corpus vs. child direct speech (CHILDES) corpus

4.4 SCF Comparison

A subset of the constructed lexicons were compared for subcategorization similarities between the BNC corpus and CHILDES corpus. To obtain reliable results, we restricted our scope to 93 verbs—all those for which the total number of sentences analysed for SCFs was greater than 50 in both corpora, and which were thus less likely to be affected by sparse data problems during SCF acquisition. The SCF lexicons for these verbs were also contrasted against the gold standard described earlier in section 4.2.

The average number of SCFs taken by studied verbs in the two corpora proved quite similar, although verbs in BNC took on average a larger number of SCFs (19) than those in CHILDES (15). However, we found that most verbs (regardless of their frequency in the corpora) showed substantially richer subcategorization behaviour in the BNC than

	BNC	CHILDES
Precision	51.41	52.21
Recall	28.57	24.36
F Measure	36.73	33.22

Figure 2: Precision, recall and F Measure of CHILDES lexicon and BNC lexicon with respect to COMLEX-ANLT combined gold standard.

in CHILDES. A total of 80 frame types were hypothesised for the 93 studied verbs in the BNC, while 68 were hypothesised in CHILDES. The intersection between these frames in the corpora was not large (0.61).

To establish whether this difference was due to one lexicon being considerably less accurate than the other, we compared the SCFs in both lexicons against the gold standard. The results listed in Figure 2 show that the BNC lexicon had a slightly higher F measure than CHILDES: 36.7 vs. 33.2.[1] This was only due to the better recall of BNC (+4.21% compared with CHILDES), as CHILDES had a better precision than BNC (+0.80%). The differences in precision and recall—although fairly small—can be largely explained by the nature of SCFs in the two corpora. The smaller number of frames proposed in CHILDES were less complex and thus easier for the system to detect correctly, while the more varied SCFs in the BNC were more complex and also more challenging for the system.

Indeed the distributions of SCFs in the two corpora appeared fairly different. As shown in Figure 3, there was only a weak rank correlation between the frames in the distributions (0.46). The Kullback-Leibler distance denotes a low degree of correlation (1.0) and the results with other measures of distributional similarity are equally unimpressive (e.g. the cross entropy is 2.7).

Our thorough qualitative analysis of SCF differences in the two corpora revealed reasons for these differences. The most basic SCFs (e.g. intransitive and simple NP and PP frames; which describe e.g. *he slept*, *he ate an apple* and *he put the book on the table*) appeared equally frequently in both corpora. However, a large number of more complex frames

[1]Note that these figures are not impressive as performance figures, largely due to the fact that the gold standard was not fully accurate as it was obtained from dictionaries rather than from the corpus data. It was also too ambitious considering the size of the corpus data used in our experiments and the zipfian nature of the SCF distributions (i.e. many SCFs listed in large dictionaries were simply missing in the data, as the low recall indicates). However, the gold standard was adequate for the purpose of these experiments.

	CHILDES vs. BNC
KL distance	1.022
JS divergence	0.083
cross entropy	2.698
skew divergence	0.533
rank correlation	0.463
intersection	0.608

Figure 3: Average similarity values

were either very low in frequency or altogether absent in CHILDES. For example, the verb *hear* appeared only in the following kind of constructions in CHILDES:

1. *I heard you*
2. *I heard*
3. *I heard that you came*

while in BNC it also appeared in the following kind of constructions:

1. *I heard it from him*
2. *Can you hear this out?*
3. *Did you hear whether he will come?*
4. *I heard him singing*

Several types of SCFs were poorly covered or largely absent in CHILDES. Many of these were frames involving sentential and predicative complementation (e.g. *I caught him stealing*, *he forgot what to do*, *I helped him to dress*) and verb-particle constructions (*I got him up from the bed*, *he came out poor*, *he looked it up*). Also a large number of adjectival frames were missing (e.g. *I remembered him as stupid*, *It dropped low*). On the other hand, frames involving prepositional or nominal complementation were covered fairly well in CHILDES (e.g. *I will get it from him*, *she built me this castle*).

While the SCF differences seem fairly big, they are not altogether arbitrary. Rather, they seem somewhat correlated with different verb senses and SCFs typically permitted by the senses. To gain a better understanding to this, we looked into Levin's taxonomy (Levin, 1993) which divides English verbs into different classes on the basis of their shared meaning components and similar syntactic (mostly subcategorization) behaviour. For example, in Levin's resource, verbs such as *fly, move, walk, run* and *travel* belong to the same class as they not

only share a similar meaning but also take similar SCFs.

When we compared for some of our test verbs the SCFs in the two corpora to those listed in Levin, we noticed that many of the SCFs absent in CHILDES and listed in the BNC and were just syntactically more complex manifestations of the same verb sense as that described by the CHILDES SCFs. For example, verb senses that take multiple sentential and predicative complements in Levin take just a smaller range of those SCFs in CHILDES than in BNC. However, some SCFs in BNC describe verb senses which were altogether absent in CHILDES. After a closer look, many of these senses proved to be extended senses of those exemplified in CHILDES.

In the light of this small scale investigation with Levin classes, it seems to us that to gain a better understanding of SCF differences in adult and CDS speech and the role of SCFs in language acquisition, it would be useful, in the future, to investigate to what extent SCF learning is mediated by the sense of the predicate and its membership in classes such as Levin's.

5 Observations

Some prevailing theories of language acquisition (e.g. that of Landau & Gleitman (1985)) suggest that verb SCFs provide convergent evidence on the meaning of a verb. These theories rely on the assumption that the frames provided in a child's input are adequately diverse to support learning. Meanwhile, Snow (1986) suggests that CDS plays an important rôle in the facilitation of acquisition. If Snow and Landau & Gleitman are both correct then we would perhaps hope to find that CDS is diverse in terms of its SCFs.

This appears to conflict with earlier small-scale empirical studies (e.g. (Snow, 1986)) which suggest that while CDS is quite complex (displaying, for example, the full range of conventional indirectness) it is syntactically much simpler than speech between adults. Our empirical results obtained from automatic SCF analysis of large-scale data[2] show conclusively that CDS is not only significantly simpler but also syntactically very different than speech between adults. Perhaps then, the rôle of CDS is to encourage the acquisition of simple frames, providing a basis from which more complex frames may be developed.

The fact that there is little correlation between the SCFs in two corpora is a little surprising as one might expect CDS to contain a subset of adult speech's SCFs. However, as our small scale experiment with Levin classes suggests, the SCFs seem nevertheless correlated via verb senses. While this issue requires further investigation, it is important to also note that some CHILDES SCFs absent in BNC may not be altogether absent in adult speech. Due to the Zipf-like nature of the SCF data, they may just occur in adult speech with a very low frequency. If this turns out to be the case after further larger scale experiments, it would indicate that most CDS SCFs are indeed a subset of those in adult speech but the frequencies of the SCF in the two corpora differ substantially.

Our results may also support Valian's (1990) findings that 4% of parental replies to children are ungrammatical, and 16% grammatical but not fully acceptable (examples from our CDS corpus include "play this together?", "another one missing."). Such utterances explain at least partly why there are SCFs present in the CHILDES lexicon that are missing from the BNC. Valian also found that adults tend to reply to children using an utterance which is lexically and structurally similar to the child's sentence (5% verbatim, 30% structurally similar). Since child speech at 2;7yrs (the average age of child subject in our CDS corpus) is usually simpler than adult speech ((Nice, 1925) and (Brown, 1973)) such repetition could help to boost the relative frequency of simpler frames in the CHILDES lexicon.

[2] We will make our data publicly available via the web.

References

B Boguraev *et al.* 1987. The derivation of a grammatically-indexed lexicon from the longman dictionary of contemporary english. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 193–200.

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

E Briscoe and J Carroll. 2002. Robust accurate statistical annotation of general text. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

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

M Demetras. 1989a. Changes in parents' conversational responses: a function of grammatical development. In *ASHA, St Louis*.

M Demetras. 1989b. Working parents' conversational responses to their two-year-old sons. University of Arizona.

C Fisher *et al.* 1994. When it is better to re-

ceive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1–4):333–375, April.

R Grishman *et al.* 1994. Comlex syntax: building a computational lexicon. In *International Conference on Computational Linguistics*, pages 268–272.

R Higginson. 1985. Fixing-assimilation in language acquisition. unpublished doctoral dissertation, Washington State University.

A Korhonen and Y Krymolowski. 2002. On the robustness of entropy-based similarity measures in evalution of subcategorization acquisition systems. In *Sixth Conference on Natural Language Learning*, pages 91–97, Taipei, Taiwan.

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

B Landau and L Gleitman. 1985. *Language and Experience: evidence from the blind child*. Harvard University Press, Cambridge, Mass.

G Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

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

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

B MacWhinney, 1995. *The CHILDES project: Tools for analysing talk*. Lawerance ErlBaum Associates, Hillsdale, NJ, second edition.

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

M Nice. 1925. Length of sentences as a criterion of child's progress in speech. *Journal of Educational Psychology*, 16:370–9.

J Pine. 1994. The language of primary caregivers. In C. Gallaway and B. Richards, editors, *Input interaction and language acquisition*, pages 13–37. Cambridge University Press, Cambridge.

K Post. 1992. The language learning environment of laterborns in a rural Florida community. unpublished doctoral dissertation, Harvard University.

D Roland and D Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *COLING-ACL*, pages 1117–1121.

D Roland *et al.* 2000. Verb subcategorization frequency differences between business-news and balanced corpora. In *ACL Workshop on Comparing Corpora*, pages 28–34.

J Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K Nelson, editor, *Children's Language*, volume 4. Lawrence Erlbaum Associates, Hillsdale, NJ.

C Snow. 1986. Conversations with children. In P Fletcher and M Garman, editors, *Language Acquisition*, pages 363–375. Cambridge University Press, New York, 2nd edition.

P Suppes. 1974. The semantics of children's language. *American Psychologist*, 29:103–114.

V Valian. 1990. Logical and psychological constraints on the acquisition of syntax. In L Frazier and J Villiers, editors, *Language Processing and Language Acquisition*, pages 119–145. Dordrecht, Kluwer.

A Warren-Leubecker. 1982. Sex differences in speech to children. unpublished doctoral dissertation, Georgia Institute of Technology.