C E N T E R F O R R E S E A R C H I N L A N G U A G E

November 1999 **Vol. 11, No. 7**

The Newsletter of the Center for Research in Language, University of California, San Diego, La Jolla CA 92093
Tel: (858) 534-2536 • E-mail: info@crl.ucsd.edu • WWW: http://www.crl.ucsd.edu/newsletter

• • •

FEATURE ARTICLE

Could Sarah Read the Wall Street Journal?

Ezra Van Everbroeck (ezra@ucsd.edu)
Department of Linguistics
University of California, San Diego

EDITOR'S NOTE

This newsletter is produced and distributed by the **CENTER FOR RESEARCH IN LANGUAGE**, a research center at the University of California, San Diego that unites the efforts of fields such as Cognitive Science, Linguistics, Psychology, Computer Science, Sociology, and Philosophy, all who share an interest in language. We feature papers related to language and cognition distributed via the World Wide Web) and welcome response from friends and colleagues at UCSD as well as other institutions. Please visit our web site: “http://www.crl.ucsd.edu”.

SUBSCRIPTION INFORMATION

If you know of others who would be interested in receiving the newsletter, you may add them to our email subscription list by sending an email to majordomo@crl.ucsd.edu with the line "`subscribe newsletter <email-address> `" in the body of the message (e.g. `subscribe newsletter jdoe@ucsd.edu`).

Please forward correspondence to:

Cristina Saccuman and Kathryn Kohnert, Editors
Center for Research in Language, 0526
9500 Gilman Drive, University of California, San Diego 92093-0526
Telephone: (858) 534-2536 • E-mail: editor@crl.ucsd.edu

Back issues of this newsletter are available from CRL in hard copy as well as soft copy form. Papers featured in previous issues include the following:

Analogic and Metaphoric Mapping in Blended Spaces: Menendez Brothers Virus
Seana Coulson
Cognitive Science, UCSD
Vol. 9, No. 1, February 1995

In Search of the Statistical Brain
Javier Movellan
Department of Cognitive Science, UCSD
Vol. 9, No. 2, March 1995

Connectionist Modeling of the Fast Mapping Phenomenon
Jeanne Milostan
Computer Science and Engineering, UCSD Vol. 9, No. 3, July 1995

Representing the Structure of a Simple Context-Free Language in a Recurrent Neural Network: A Dynamical Systems Approach
Paul Rodriguez Department of Cognitive Science, UCSD Vol. 10, No. 1, October 1995

A Brain Potential Whose Latency Indexes the Length and Frequency of Words
Jonathan W. King
Cognitive Science, UCSD
Marta Kutas
Cognitive Science and Neurosciences, UCSD
Vol. 10, No. 2, November 1995

Bilingual Memory: A Re-Revised Version of the Hierarchical Model of Bilingual Memory
Roberto R. Heredia
Center for Research in Language, La Jolla, CA
Vol. 10, No. 3, January 1996

Development in a Connectionist Framework: Rethinking the Nature-Nurture Debate
Kim Plunkett
Oxford University
Vol. 10, No. 4, February 1996

Rapid Word Learning by 15-Month-Olds under Tightly Controlled Conditions
Graham Schafer and Kim Plunkett
Experimental Psychology, Oxford University
Vol. 10, No. 5, March 1996

Learning and the Emergence of Coordinated Communication
Michael Oliphant and John Batali
Department of Cognitive Science, UCSD
Vol. 11, No. 1, February, 1997

Contexts That Pack a Punch: Lexical Class Priming of Picture Naming
Kara Federmeier and Elizabeth Bates
Cognitive Science, UCSD
Vol. 11, No. 2, April, 1997

Lexicons in Contact: A Neural Network Model of Language Change
Lucy Hadden
Department of Cognitive Science, UCSD
Vol. 11, No. 3, January, 1998

On the Compatibility of CogLexicons in Contact: A Neural Network Model of Language Change
Mark Collier
Department of Philosophy, UCSD
Vol. 11, No. 4, June, 1998

Analyzing Semantic Processing Using Event-Related Brain Potentials
Jenny Shao
Department of Speech Pathology, Northwestern University
Helen Neville
Department of Psychology, University of Oregon
Vol. 11, No. 5, December 1998

Blending and Your Bank Account: Conceptual Blending in ATM Design
Barbara E. Holder
Department of Cognitive Science, UCSD
Vol. 11, No. 6, April 1999

Could Sarah Read the Wall Street Journal?

Ezra Van Everbroeck (ezra@ucsd.edu)
Department of Linguistics
University of California, San Diego

Abstract

In this paper I compare the semantic and syntactic properties of 2,000 verbs from two very different types of text: half of the corpus came from Child-Directed Speech (CDS) to Sarah (Brown 1973), while the other half was taken from the business section of the Wall Street Journal (WSJ). Each verb was tagged with its syntactic subcategorization frame of complements and adjuncts, and it was also noted to which of Vendler's (1967) four conceptual categories it belonged. Finally, the voice, polarity and mood of each verb were established. The comparison of verbs across the two texts reveals semantic similarities, although the verbs themselves tend to appear in different syntactic constructions. Interestingly, the Child-Directed Speech text is, in some linguistic areas, more complex than its Wall Street Journal counterpart.

1. Introduction

How do semantic and syntactic properties of verbs in texts drawn from different sources match up? Theoretical linguists of many persuasions have discussed these properties in great detail (e.g. Dowty 1979; Jackendoff 1991; Langacker 1991; Goldberg 1995). Cognitive psychologists have also contributed a number of valuable developmental and experimental studies (e.g. Marantz 1982; Fisher *et al.* 1991; Gropen *et al*. 1991). In recent years, connectionist modelers have become involved as well, trying to make their simulations sensitive to the properties of the different classes of words in the input languages of the models (e.g. Elman 1993). Their wildly different goals and methods notwithstanding, one type of data which is (almost) completely absent in all these studies is statistical information about the properties of verbs as they occur in real language. This paper is a preliminary step in filling this hole.

I analyzed 2,000 verbs as found in English Child-Directed Speech (CDS) and articles from the Wall Street Journal (WSJ), and coded each of them for a number of characteristics. A semantic classification system was used, the subcategorization frame was established, and notes were made on the mood, voice, and polarity of the verbs. Given the purely descriptive aim of the investigation, I will make no attempt here to use the results for any theoretical or experimental purposes, though such extensions are obviously both feasible and desirable. For example, linguists can benefit from the data presented below by using it to evaluate different theoretical proposals. Some issues, like the relevance of Vendler's (1967) semantic verb classification, are a step closer to being settled, while new fuel is also provided for further discussions. From a different perspective, developmental psychologists may be surprised to learn not only that there are some linguistic areas in which CDS is almost indistinguishable from what can be found in the business section of Wall Street Journal, but also that, in some linguistic respects, the former can be more complex than the latter. For example, CDS contains more modal verbs, more negative polarity sentences, and a greater number of questions than the WSJ text. Connectionist researchers interested in modeling language acquisition and development should be aware of the linguistic properties of "real world" corpora if they are to implement computational models in a viable manner (Rispoli 1999). Obviously, knowing the relative frequency of occurrence of different verb types with respect to their semantic properties, as well as the syntactic patterns in which these verbs occur, are required if one wants to model realistic language input.

I will first discuss in some detail the various characteristics used for tagging the corpus. I will then present the overall results obtained. The conclusion deals with the issue of how the coding methods described in this paper can be improved.

2. Methodology

2.1 The Corpus

The two texts chosen for the corpus are a number of fragments of Child-Directed Speech (CDS) as well as a number of articles from the Wall Street Journal (WSJ). Dimensions along which these texts differ include spoken vs. written language, informal vs. formal language, conversation vs. exposition, the amount of presumed background knowledge in the hearer/reader, and the nature of the topics discussed. Though it may be possible to find registers which are further apart than these two, they are so intuitively dissimilar that they appear well-suited for finding linguistic differences between them. I will discuss each text in greater detail in the following paragraphs.

The fragments of CDS were taken from the CHILDES corpus (MacWhinney 1995; Sokolov & Snow 1994), from Roger Brown's (1973) description of the interaction between the child Sarah and her working-class parents (or other adults present in the room, including the investigators). No attempt was made on my part to distinguish between utterances directed at Sarah, and those spoken by one adult to another one. Not only were the latter quite rare, but Sarah heard them too, so they were an integral part of the linguistic input which she received. In order to see whether there would be any change in the CDS spoken as a function of Sarah's development, I selected four different fragments of the Sarah corpus. The fragments were from the recordings when Sarah was 2;3, 3;0, 4;0, and 5;0 years old respectively. From each fragment, the first 250 verb tokens spoken by adults were analyzed for their semantic and syntactic properties. However, because of their limited relevance for this study, I ignored tag questions, incomplete sentences (in that they did not contain a verb), incomprehensible utterances as well as ungrammatical sentences when the adult was simply repeating what Sarah had just said to him or her. The presence of modal verbs such as *will, have to* or *may* was noted (and they counted towards the number of 250), though they were not analyzed further (see below). Here is a representative excerpt from the text, with the interaction between Sarah (SAR) and her mother (MOT). Italics indicate the verbs which were coded.

> *MOT: Sarah # *don't touch* it.
> *SAR: I want xx ribbon on mine.
> *MOT: you *want* ribbons on yours?
> *MOT: all your ribbon *is* down (a)t the beach.
> *MOT: Mommy *didn't bring* any ribbon home.
> *SAR: you got yours [= microphone]?
> *MOT: yeah.

The text from the Wall Street Journal (WSJ) consisted of a number of consecutive articles printed in the November 2, 1989 edition. Though there were a few very brief articles (mainly dealing with people being named to certain positions, or people resigning), most of the newspaper articles contained more than 20 (often very long) sentences and discussed various issues related to business (e.g. lawsuits, factory openings, and an analysis of the computer market). In order to be able to do within-text comparisons, I also divided the total selection of 1,000 verbs into four fragments with 250 verbs each. As one might expect from a newspaper like the Wall Street Journal, all sentences containing verbs were both comprehensible and complete, and were thus coded. The italics in the following excerpt indicate which verbs were tagged.

> Yields on money market mutual funds *continued* to *slide*, amid signs that portfolio managers *expect* further declines in interest rates. The average seven day compound yield of the 400 taxable funds *tracked* by IBC/Donoghue's Money Fund Report *eased* a fraction of a percentage point to 8.45% from 8.47% for the week *ended* Tuesday. Compound yields *assume* reinvestment of dividends and that the current yield *continues* for a year. Average maturity of the funds' investments *lengthened* by a day to 41 days, the longest since early August, according to Donoghue's. Longer maturities *are thought* to *indicate* declining interest rates because they *permit* portfolio managers to *retain* relatively higher rates for a longer period.

2.2 The Classifications

Not only were verb properties analyzed, but the actual coding also took into account the entire clause in which each verb appeared. This was necessary to establish the (syntactic) subcategorization frame with which the verbs occurred as well as to determine the semantic class of each verb. Also, the form of the main verb by itself does not always unambiguously signal the voice or mood of an English clause or sentence. The final classification scheme used for each verb contained six fields. The structure of these six

classification fields as well as a brief description of each is presented below.

1. The verb itself (in its uninflected form)
2. The Vendler classification for the verb (empty for modals)
3. A list of complements or adjuncts appearing with the verb (possibly empty)
4. The voice of the verb (default: active)
5. The polarity of the verb (default: affirmative)
6. The mood of the sentence (default: indicative)

In accordance with this classification structure, the verb *eat* in the sentence *Wasn't he eaten by a crocodile last week?* would have been tagged as follows:[1]

eat; accomplishment; by-phrase time-adjunct; passive; negative; interrogative

A more usual sentence such as *I like bananas* would have received a much simpler tagging:

like; state; direct-object

The information stored in fields 1, 4, 5 and 6 is relatively straightforward. However, the classifications used for fields 2 and 3 need further clarification. Thus, in the following section I will describe both the Vendler classification system (2) as well as the subcategorization system (3) used in the current verb coding system.

2.2.1 The Vendler Classification

In his seminal paper, Zeno Vendler (1967) proposed that English verbs could be grouped into four distinct semantic categories, each with its own characteristics and entailments. A first dimension along which verbs differ is whether they denote tenses which are continuous (e.g. *run, draw a circle*) or finite (e.g. *know, recognize*). The former could again be divided into two categories, depending on whether the process has a natural climax or end point (i.e. accomplishments) or not (i.e. activities). Verbs which do not have continuous tenses also fall into two categories, but the crucial point this time is whether their meaning contains an inherent aspect of duration (i.e. states) or not (i.e. achievements). The following examples serve to illustrate each of these four classes.

> Accomplishments: to build a house, to play a game of basketball, to take a picture, to write a paper or to say a word. All these verb phrases refer to processes which take some time, but which end when the goal has been reached.
>
> Activities: to run, to play, to climb, to sing or to cry. As with the accomplishments, the activities take up time, however, unlike the previous category, these actions could theoretically go on forever.
>
> States: to know something, to like someone, to be human, or to own a country. States describe (almost) immutable situations in the world, so they are similar to activities in that the duration of the process is unbounded, but unlike the activities in that they do not imply a volitional agent undertaking the process.
>
> Achievements: to reach the top, to win a championship, to smell the gas or to die. Unlike all the other classes, achievements do not inherently take time; rather they denote instantaneous changes of situations and, in that regard, they resemble accomplishments quite closely.

As the occurrence of *play* in both the list of accomplishments and the list of activities shows, a single English verb can belong to more than one of these classes, depending on its exact meaning and the other words which appear in the same clause (e.g., see Vendler, 1967; Van Valin & Wilkins 1993). For example, Van Valin & Wilkins (1993) point out that *remember* functions as an achievement in *She suddenly remembered the towel she had left at the beach*, as an activity in *He consciously remembered the faces of all the people he had seen at the conference*, and as a state in *I remember my first day in grad school pretty vividly*.[2] In his paper, Vendler discusses some linguistic tests to determine which class a verb belongs to (e.g. whether the verb can have continuous tense marking), but it is David

[1] The negative marking for *eat* here may come as a surprise, but it is important to keep in mind that the tagging only looked at the verb form for this field. Also, the fact that it is not possible to add another negation marker to the verb in such a sentence to achieve "true" negative polarity suggests that the *not* which is present serves part of its normal function.

[2] It is also possible to get an accomplishment reading of *remember* but one needs to construct a special sentence in which words are added to stress duration and to avoid the impression of a sudden change of state, e.g. *It took him a long time to gradually remember my name*, which admittedly sounds quite unnatural.

Criterion	States	Activities	Achievements	Accomplish-ments
Verb occurs in progressive		Y		Y
Verb occurs as imperative		Y	Y	Y
Verb occurs with *carefully, attentively*		Y		Y
Verb has habitual reading in simple present		Y	Y	Y
take an hour to Verb; Verb *in an hour*			Y	Y
Verb *for an hour; spend an hour* Verb-*ing*	Y	Y		Y
x is Verb-*ing* entails *x has* Verb-*ed*	N/A	Y	N/A	
Verb occurs as complement of *stop*	Y	Y		Y
Verb occurs as complement of *finish*				Y
Verb is ambiguous with *almost*				Y
Verb occurs as complement of *persuade*		Y		Y
Verb *in an hour; take an hour to* Verb			Y	Y

Table 1. Some of the criteria used to determine the Vendler class of the verbs in the corpus. A 'Y' means that the verb can occur in that construction without causing a semantically odd reading; N/A means that the test does not apply reliably to the verbs of this class (compare Vendler 1967: 60, Table 1).

Dowty (1979) who has presented a comprehensive analysis of the conceptual and linguistic properties of the four classes. His criteria form the basis for the way the 2,000 verbs in the present corpus were tagged, and some of the more important ones are summarized in Table 1.

2.2.2 The 'Subcategorization' Classification

Next to the semantic classification, an extra field was reserved for more syntactic information. Each verb in the corpus was tagged with the complements and adjuncts with which it appeared; the presence of a subject was not coded, as it is obligatory in English. Given that adjuncts are not usually thought of as being subcategorized for in the lexicon, the term 'subcategorization (frame)' is somewhat misleading, but I will continue to use it with this extended meaning throughout the paper. The presence of the following elements was coded:

direct object
indirect object
complement *that*-clause (e.g. *he thinks* ***that*** ...)
complement *if*-clause (e.g. *he asked* ***if*** ...)
complement quote (e.g. *he said* ***"..."***)
complement infinitival clause (e.g. *he wants* ***to*** ...)
complement *ing*-clause (e.g. *he tried* ***... -ing***)
predicate (e.g. *he was* ***happy***)
agentive *by*-phrase in passive
prepositional phrase
adverbial, PP or clause expressing reason for action
adverbial, PP or clause expressing moment/duration of action
adverbial or PP expressing location of action
adverbial or PP expressing manner of action
idiom

Word order was not coded but, as noted before, aspect, mood and polarity were included.

3. Results and Discussion

3.1 Types of Verbs

A first, basic question one might want to ask about the corpus is how many different verb *types* it contains. Table 2 below summarizes the data for each of the eight corpus fragments (four CDS; four WSJ)

DIFFERENT VERB TYPES

	Child-Directed Speech					Wall Street Journal					Total
age:	1 2:3	2 3:0	3 4:0	4 5:0	Sub-total	1	2	3	4	Sub-total	
verbs	40	74	80	75	149	141	131	138	140	387	483

Table 2

MODAL VERB TOKENS

	Child-Directed Speech					Wall Street Journal					Total
	1	2	3	4	Sub-Total	1	2	3	4	Sub-Total	
modals	13	32	30	33	10.80%	15	35	20	16	8.60%	9.70%

Table 3

of 250 verb tokens produced. The fifth column within each text type contains the total number of different verbs (tokens) for either CDS or WSJ, respectively. The final Total column shows the same information calculated over the entire corpus at once.

There are at least three points worth noting about the table. First, the WSJ text contains more than twice as many distinct verbs as the CDS text (387 > 149). Secondly, there is basically a doubling of the number of verbs found in the CDS text for Sarah at age 2 to Sarah at age 3 (CDS corpus 2 relative to CDS corpus 1), but the number stays almost the same afterwards (a point I will return to below). Finally, there is a significant amount of overlap between the verbs in the two parts of the corpus (there are only 483 different types in the overall corpus, many fewer than the 536 (149 + 387) which one would have expected to find if there was no overlap at all). In general, then, these results support the intuitively plausible hypotheses that 1) the WSJ is more difficult (because it is more varied) than the input Sarah (or any other young child) receives and that 2) there is a core of English verbs which appears in all possible registers. Despite the existence of such a core (and not withstanding substantial lexical and conceptual knowledge), the numbers presented here do make it seem unlikely that Sarah — at least at age 2 — would have been able to understand all of the Wall Street Journal.

Another way of looking at the general properties of the verbs in the corpus, however, is to determine how many of them were 'real' verbs, as opposed to modals. Table 3 below provides the results for this question. (The first four columns contain the absolute number of modals found in each of the fragments, while the three Total columns show the percentages relative to the two texts, and the entire corpus, respectively.) The numbers for the 'real' verbs are not given, because they are simply the complement of the modal sums.

Perhaps somewhat surprising is the result that the CDS text contains a greater percentage of modals than the WSJ, especially if one takes into account that the first CDS fragment is again distinct from the other ones. On the other hand, if one knows that the most frequent modals in the Child-Directed Speech text are instances of *can* (21) — as in *you can't do* ... — and *have to* (22) — as in *you have to* … —, then this result becomes much more understandable. That is, the adults are telling Sarah what she can and cannot do, which is not exactly the kind of linguistic activity one would expect to find in a newspaper like the Wall Street Journal.

3.2 Polarity, Voice, Mood

The second set of data concern the numbers found for the clausal categories of polarity and voice, and for the sentential property of mood. Table 4 summarizes the results.

	Child-Directed Speech					Wall Street Journal					Total
	1	2	3	4	Sub-Total	1	2	3	4	Sub-Total	
negatives	21	30	26	16	9.30%	1	9	2	6	1.80%	5.55%
passives	0	0	0	0	0.00%	29	27	35	24	11.50%	5.75%
questions	128	58	56	53	29.50%	0	0	0	1	0.10%	14.80%
orders	20	38	24	20	10.20%	0	0	0	1	0.10%	5.15%

Table 4

As far as polarity and mood are concerned, we find a continuation of the trend just discussed. Specifically, the CDS text is considerably more complicated than the WSJ, with more than five times as many negated verbs, and almost 40% of its sentences with a marked, non-declarative mood (as opposed to less than 1% for the WSJ). The two are however not closely correlated: only 12 of the 102 imperatives are negative, and only 8 of the 295 questions are in the negative form. The huge number of questions is mostly the result of the fragment for Sarah at age 2. We will see below that this is the result of a single construction occurring very frequently in the CDS input at this age.

The results for the passive voice are exactly opposite to this trend. Here, the WSJ text features increased complexity, with more than 10% of its verbs in the passive construction.[3] The complete absence of passives in the CDS corpus is not a coding mistake. There are actually only a few passives in the Sarah corpus at large (Brown 1973), but none in the fragments analyzed here. What these numbers suggest, then, is that a possible reason why passives are acquired quite late by children learning English may not be related so much to some perceived inherent difficulty with the active-passive transformation, but rather to the infrequency of the passive construction in the child's input. This is consistent with the findings of Demuth (1990), who found that children learning Sesotho start producing passive constructions much earlier than children learning English do. However, passives are also used a lot more frequently by adult speakers of Sesotho.

3.3 Vendler Classification

In Table 5 below, I present the frequency results for the Vendler classification as applied to the verbs in the corpus.

The one number which stands out here is the frequency of state verbs in the first fragment of the Child-Directed Speech text. As we have seen already that this fragment is unusual in other respects as well. It is worth checking what happens when we discount the data in this fragment and calculate the relative frequencies of the four verb categories using the data from the other three fragments. Table 6 below provides the new results.

Although there are differences between the two parts of the corpus, these differences are not dramatic. The largest difference is for the state verbs at slightly less than 10%, while the percentage of achievements is almost identical in both texts. So, in comparison to some of the measures described earlier, the Vendler classification does not appear to be very useful in distinguishing between different registers of language. A closer look at the two corpora reveals that there are four semantic-domain classes of verbs for which there are larger differences; existence, speech, size, and showing. It turns out that all four can be quite easily accounted for by considering particular discourse

[3] Though it does not show up in any of the tables presented here, another major difference in complexity between the Child-Directed Speech text and the WSJ is in the length and internal structure of their noun phrases: in the former, they tend to be very short (e.g. pronouns), but in the WSJ they are often longer than five words.

VENDLER CLASSIFICATION

	Child-Directed Speech					Wall Street Journal					Total
	1	2	3	4	Sub-Total	1	2	3	4	Sub-Total	
accomp	23	66	89	42	22.00%	55	79	87	82	30.30%	26.15%
achieve	17	25	35	49	12.60%	66	40	34	23	16.30%	14.45%
states	187	78	73	99	43.70%	62	62	51	64	23.90%	33.80%
activities	9	49	23	28	10.90%	52	34	58	65	20.90%	15.90%

Table 5

VENDLER CLASSIFICATION (WITHOUT FIRST CDS PERIOD)

	Child-Directed Speech					Wall Street Journal					Total
	1	2	3	4	Sub-Total	1	2	3	4	Sub-Total	
accomp	0	66	89	42	26.27%	55	79	87	82	30.30%	28.28%
achieve	0	25	35	49	14.53%	66	40	34	23	16.30%	15.42%
states	0	78	73	99	33.33%	62	62	51	64	23.90%	28.62%
activities	0	49	23	28	13.33%	52	34	58	65	20.90%	17.12%

Table 6

topics. For the latter three classes, the WSJ text has many more items, but this is what one would expect if there were a need to report on what important business people said (speech: *said, announced, told*), how the stocks of companies are doing (size: *increase, decrease, grow*) and what various economic indicators show (showing: *indicate, register, show*). While these same classes could also be used in a more informal and conversational register, they are not as inherently important for it as they are for a newspaper which survives only because it provides this kind of information. The overabundance of *existence* verbs in the CDS text, and especially its first fragment when Sara was two years old, finally allows us to determine what has been skewing the data all along. It is the existence verb *be* as used in the questions *What is this?* or *Where is your nose?* . These types of questions are extremely frequent in the first CDS fragment. One plausible explanation for this phenomenon is that Sarah had not yet begun the productive vocabulary burst; hence, the desire on the part of the adults to teach her about the meanings of (new) words. The fragments which I have looked at also show that at age 2, Sarah was definitely not as talkative as she was as at age 5, another possible reason for why the adults in her environment were asking her questions.

Another point worth noting is that there is a fairly large amount of internal variation between the fragments of even a single text. This suggests strongly that the precise frequencies are still subject to considerable change if the corpus were to be enlarged. (The overall similarity between the two texts, though, indicates that there is some validity to the frequencies obtained here.) Hence, whether activity verbs are really more common than achievement verbs should be left as an open question, but it is probably the case that accomplishment and state verbs make up more than half of the verbs in various English texts.

SUBCATEGORIZATION RESULTS

	Child-Directed Speech					Wall Street Journal					Total
	1	2	3	4	Sub-Total	1	2	3	4	Sub-Total	
direct object	34	106	110	74	32.40%	106	85	98	81	37.00%	34.70%
indirect object	5	13	9	6	3.30%	8	6	11	4	2.90%	3.10%
that-clause	4	7	8	9	2.80%	21	29	25	27	10.20%	6.50%
if-clause	1	0	0	0	0.10%	0	0	0	1	0.10%	0.10%
quote	0	0	0	0	0.00%	6	7	4	5	2.20%	1.10%
to infinitive	9	12	6	8	3.50%	9	13	22	12	5.60%	4.55%
-ing clause	0	2	0	0	0.20%	0	3	1	1	0.50%	0.35%
predicate	162	32	45	44	28.30%	34	32	32	30	12.80%	20.55%
by phrase	0	0	0	0	0.00%	3	3	10	2	1.80%	0.90%
prep phrase	4	21	12	9	4.60%	27	32	25	33	11.70%	8.15%
causal phrase	1	7	6	1	1.50%	6	9	7	4	2.60%	2.05%
time phrase	3	8	14	3	2.80%	34	31	23	33	12.10%	7.45%
location phrase	25	35	39	34	13.30%	16	8	11	8	4.30%	8.80%
manner phrase	3	1	2	10	1.60%	14	7	9	30	6.00%	3.80%
idiom	1	4	2	2	0.90%	2	0	1	0	0.30%	0.60%
Sum	252	248	253	200	95.30%	286	265	279	271	110.10%	102.70%

Table 7

It is interesting that all four classes of verbs have reasonably similar frequencies. There is nothing inherent in the Vendler classification which forces this to be the case, so the numbers in Table 6 could be used to defend the usefulness of the classification against reductionist alternatives: if two classes were to be collapsed, we would lose specific information about at least 15% of the verbs in the corpus.

3.4 Subcategorization Classification

The syntactic results for the subcategorization frames are presented in Table 7. The format of the table is again similar to the ones previously discussed, although I have added a final row which contains the sum of all the numbers in the rows above it. The numbers in this row indicate that the clauses of the WSJ have greater syntactic complexity than the ones in the CDS text, even if we limit ourselves to just counting the number of constituents.

The results in Table 7 support what we have previously noted. Specifically, there is a large degree of similarity between the CDS text and the WSJ text. However, this generalization does not hold for *that*-complements (which accompany the Speech verbs in the WSJ text), and predicates (which accompany the Existence verbs in the CDS text). As for the adjuncts, we find an irregular pattern in that although prepositional phrases, adverbs of time, and adverbs of manner are noticeably more frequent in the WSJ, adverbs of location are dominant in the CDS text. What may be at the root of this mismatch is that the conversational setting of the Sarah fragments

supports a lot of pointing at things, and the use of deictic words like *here* and *there*. This is obviously not the case for the WSJ text, but it does have to report on when events took place: e.g. when and how stock fluctuations occurred.

4. Conclusion

In this paper, I have compared the semantic and syntactic properties of 2,000 verb tokens found in two very different types of English texts. The results of the investigation revealed that there are large differences between the two texts in some areas (mood, polarity, voice, and the number of different verb used.) Interestingly, however, the analysis also revealed many similarities in their semantic and syntactic characteristics. The former were found to be largely parallel by a different measure, namely the four-way split of the Vendler classification. The syntactic properties were compared on the basis of the adjuncts and complements with which the verbs appeared in the corpus texts. Given these results, we may wonder what predicament this leaves Sarah in, when she is presented with a recent edition of the Wall Street Journal. The tentative conclusion seems to be that she would do a decent job - if only she knew the meaning of the words! Syntactically, the subcategorization frames in the WSJ text would not surprise her, although the length of the sentences might present problems. Semantically, she already knows the different types of verbs. So, it seems that the CDS input has prepared her reasonably well for a business career.

However, the comparison done here should be considered preliminary and could be expanded on in a number of ways. For example, increasing the raw number of verbs coded would provide more power for statistical analysis. In addition, examining texts from registers other than the two analyzed here would make it possible to sketch a picture of the entire continuum of registers, rather than two extreme endpoints. Still, the between-sample consistency is quite high. If one looks again at Tables 6 and 7, one can easily see that the three later CDS and all four WSJ texts share many linguistic properties. For example, there are about eight *that*-complements in every 250 verbs in the CDS corpus, whereas the number for the WSJ texts is consistently around 25. The small differences between the fragments of a single register hint at quite robust data. Another improvement would be to have two persons tag the same verbs independently to determine inter-coder reliability. This would obviously decrease the risk of coding errors, but it would make the verb tagging an even more time-exhaustive enterprise.[4] Concerning the method used to tag the verbs, it would have been better if tense and aspect had been coded for, as well as clause-level polarity, rather than the current verb-only criterion. Also, it would probably have been useful to store the information about the order of the constituents in each clause. Finally, coding the internal structure of both noun phrases and entire sentences (i.e. the relationships between the clauses) would allow a more precise characterization of the linguistic properties of the different parts of the corpus.

Acknowledgments

I would like to thank Liz Bates, Kathi Kohnert, Cristina Saccuman and at least one anonymous reviewer for their useful comments on an earlier version of this paper. The author retains full responsibility for any remaining bugs and mistakes, though.

REFERENCES

Bowerman, M. (1994 [1989]), "Learning a Semantic System: What role do cognitive predispositions play?", in *Language Acquisition. Core readings* (ed. Bloom), Cambridge, MA: MIT Press, 329-363.

Brown, R. (1973), *A First Language: The Early Stages*, Cambridge, MA: Harvard.

Demuth, K. (1990), "Subject, Topic and Sesotho Passive", in *Journal of Child Language* 17.1, 67-84.

Dowty, D. (1979), *Word Meaning and Montague Grammar*, Dordrecht: Reidel.

Elman, J. (1993), "Learning and development in neural networks: the importance of starting small", *Cognition* 48, 71-99.

[4] For this project, each verb was coded by hand, a process which would at least take a minute from finding the verb in the text to entering the list of its properties in a file. Verbs for which the semantic Vendler classification was not immediately obvious took several times longer to tag, because I had to consider the various criteria given in Table 1. It would be possible to automate the tagging process at least partly by having an automatic parser go over the text first and only present the human tagger with the options which seem likely.

Fisher, C., H. Gleitman, and L. Gleitman (1991), "On the Semantic Content of Subcategorization Frames", *Cognitive Psychology* 23, 331-392.

Gleitman, L. (1994 [1990]), "The Structural Sources of Verb Meanings", in *Language Acquisition. Core readings* (ed. Bloom), Cambridge, MA: MIT Press, 174-221.

Goldberg, A. (1995), *Constructions. A Construction Grammar Approach to Argument Structure*, Chicago: Chicago University Press.

Gropen, J., S. Pinker, M. Hollander, and R. Goldberg (1997), "Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure", *Cognition* 41, 153-195.

Jackendoff, R. (1991), "Parts and boundaries", *Cognition* 41, 9-45.

Langacker, R. (1991), *Foundations of Cognitive Grammar. Volume II,* Stanford: Stanford University Press.

Levin, B. and M. Rappaport Hovav (1991), "Wiping the slate clean: A lexical semantic exploration", *Cognition* 41, 123-151.

MacWhinney, B. (1995), *The CHILDES project : Computational Tools for Analyzing Talk*, Hillsdale: Lawrence Erlbaum.

Marantz, A. (1982), "On the Acquisition of Grammatical Relations", *Linguistische Berichte* 80, 32-69.

Rispoli, M. (1999), "Rethinking innateness" [review article], in *Journal of Child Language* 26, 217-225.

Sokolov, J. and C. Snow (1994), *Handbook of research in language development using CHILDES*, Hillsdale: Lawrence Erlbaum.

Van Valin, R. and D. Wilkins (1993), "Predicting Syntactic Structure from Semantic Representations", in *Advances in Role and Reference Grammar* (ed. Van Valin), Amsterdam: John Benjamins, 499-534.

Vendler, Z. (1967), *Linguistics in Philosophy*, Ithaca, NY: Cornell University Press (Chapter 4)

Appendix

The appendix contains lists of all the verbs in each of the fragments, followed by their frequency in parentheses.

Sarah (2;3): ask (3), be (170), be_to (1), begin (1), bless (1), call up (1), can (6), come (4), come_off (1), do (1), drink (2), eat (3), fix (1), get (2), give (1), go (4), going_to (2), have (1), have_to (2), hear (2), know (7), let (1), like (2), love (1), make (1), play (1), read (2), ride (1), say (4), see (4), shout (1), talk to (1), taste (1), tell (1), think (1), touch (1), use (1), want (6), will (2), write (1)

Sarah (3;0): be (35), be_to (1), bite (2), blame (1), break (1), bring (3), camp (2), can (2), chew (1), come (2), come_off (2), come_on (2), cook (1), could (1), cry (2), do (5), drink (1), dry (2), fix (1), get (20), give (3), go (7), going (2), going_to (11), guess (1), happen (2), have (9), have_to (3), hold (1), know (7), let (4), like (1), look_for (1), lose (4), love (1), make (4), miss (2), move (1), must (1), need (2), open (1), pick_up (3), play (5), push (3), put (7), put_on (3), remember (1), say (2), see (5), should (1), sing (1), sit (2), sleep (1), spend (1), spill (2), stand_up (1), take (6), take_off (2), tell (2), think (6), touch (2), turn_around (1), turn_off (1), use (1), want (11), wash (1), watch (2), watch_out (1), will (8), wipe (2), worry (1), would (2), write (5), yell (3)

Sarah (4;0): bawl (1), be (35), be_to (1), bet (1), break (2), bring (1), buy (1), can (7), clean_out (1), come (3), come_down (1), come_off (2), come_on (1), could (1), cry (1), die (1), do (12), dry_off (1), end_up (1), finish_up (1), fix (3), get (13), get_off (3), get_out (1), go (8), go_on (1), go_out (2), going_to (5), happen (5), have (16), have_to (7), hit (1), hope (1), join (1), know (6), laugh (1), leave (9), let (4), like (2), look (5), look_at (1), lose (1), make (4), mean (1), must (1), need (3), play (3), push (4), put (4), put_out (1), put_up (1), rope_up (1), say (4), scream (1), see (5), shall (1), sing (1), sit_up (1), smell (2), take (1), take_off (2), take_out (1), talk (2), taste (1), tell (4), think (6), try (1), turn (1), wait (2), wash (2), wash_off (1), watch (1), water (1), wear_out (1), wet (2), will (6), wipe_up (1), work (3), would (1), write(1)

Sarah (5;0): add (3), add_up (1), be (54), blow (1), buy (1), call (1), can (6), chop_off (1), come (3), come_off (1), come_on (1), come_to (1), could (2), do (5), fall_off (1), find_out (1), get (5), get_off (1), give (1), give_up (1), go (12), go_out (1), go_up (1), going_to (1), happen (5), have (11), have_to (10), hit (1), hurt (3), keep (1), know (6), laugh (1), learn (1), let (2), like (2), look (4), lose (1), love (3), make (7), mean (2), notice (1), pretend (2), push (1), push_around (1), put (2), put_in (1), say (3), see (13), shop (1), should (1), spell (1), start (6), stay (1), stay_on (1), suppose (1), take (1), take_off (1), talk (2), taste (1), tell (2), think (7), try (2), turn (1), turn_around (1), use (3), used_to (3), wait (4), want (5), wash (1), watch (1), will (4), work (2), would (5), write (2)

WSJ (1): act (3), announce (2), appear (2), approve (1), argue (1), assume (1), award (1), ban (1), be (32), beat (1), become (1), begin (1), blip_down (1), blip_up (1), board (1), boost (1), bring (1), can (2), capture (1), cast (1), cause (2), classify (1), complete (1), consider (2), continue (3), contract (1), cost (2), cut (1), decide (1), describe (1), diagnose (1), die (1), do_fine (1), drool_over (1), drop (1), dump (1), ease (1), eat (1), elect (1), employ (1), enter (1), exceed (1), expand (1), expect (4), expose (2), fall (1), feed (1), find (1), follow (1), gain (1), give (2), go_after (1), grow (1), hang (1), haul_out (1), have (10), hear (1), hold (3), impose (1), increase (4), indicate (1), introduce (2), invest_in (1), issue (1), jet_off (1), join (2), keep (1), know (1), lead (1), lengthen (1), lift (2), look_forward_to (1), lower (1), maintain (1), make (6), matter (1), may (1), meet (1), mix (1), name (3), offer (1), outlaw (1), oversea (1), own (1), pay (1), point_out (1), pour_in (1), pour_into (1), prove (1), race (1), raise (1), reach (1), recognize (1), record (2), register (1), regulate (1), reject (1), release (1), remain (1), replace (1), report (2), resign (1), retain (1), return (1), reward (1), rise (1), say (21), sell (2), settle_on (1), shore_up (1), should (1), show (1), show_up (1), slide (2), spend (1), squeeze_in (1), stop (2), study (3), succeed (1), support (1), suspend (1), take_place (1), talk (1), tempt (1), think (1), total (1), track (1), treat (1), try (1), underscore (1), use (4), vary (1), ventilate (1), vote (1), waive (1), watch (1), welcome (1), will (9), work (1), would (2), yield (1)

WSJ (2): add (4), announce (1), anticipate (2), appeal (2), appear (1), apply (1), approve (1), arise (1), assert (1), attach (1), attract (1), audit (1), be (22), be_able (1), be_to (1), begin (1), believe (2), benefit (1), block (1), bow_out (1), calculate (1), call (1), can (1), cause (1), change (1), close (3), collect (2), come (1), come_true (1), compare (1), compete (1), complete (3), complicate (2), consider (2), contain (1), could (6), decide (1), depend_on (1), describe (1), determine (1), disclose (1), do (1), double (1), elect (1), emerge (1), employ (1), entertain (1), estimate (1), evaluate (1), exist (1), expect (4), face (2), favor (1), file (1), find (1), follow (1), force (1), get (1), grow (1), have (3), have_to (2), head (1), help (1), hold (1), hope (2), inch_down (1), include (3), incur (1), jump (1), justify (1), lead_to (1), leave (2), leave_up (1), make (4), manufacture (1), may (3), might (1), move (1), need (1), note (2), occur (1), offer (1), open (1), order (5), own (2), pay (2), place_on (1), post (1), produce (1), propose (3), provide (1), raise (2), reach (1), receive (2), refile (1), refund (4), regard (1), relate (1), remain (1), report (2), require (2), rise (1), roll_out (1), rule (1), rule_on (1), say (24), scrap (1), seek (2), seem (2), set (4), should (1), show (1), slash (1), speed_up (1), succeed (1), suffer (1), take (1), talk (1), track_down (1), trade (1), transfer (1), turn_down (1), uphold (1), use (1), value (1), want (1), will (14), withdraw (2), work (2), worry (1), would (6)

WSJ (3): achieve (1), acquire (1), adapt (1), address (1), aid (1), allow (1), amend (1), announce (1), apply (1), approve (1), argue (1), ask (1), aspire (1), assemble (2), assist (1), be (26), become (2), believe (1), boost (1), bring (1), build (1), buy_up (1), call_for (1), can (1), change (1), channel (1), claim (1), close (1), come (2), compel (1), complain (1), complete (1), continue (1), could (5), cover (2), create (1), cut (2), decide (3), decline (1), deem (2), deny (1), design (1), deter (1), develop (3), direct (2), discourage (1), divest (1), elaborate (1), eliminate (1), enact (1), enter (2), expect (2), face (1), feel (1), file (1), force (1), forgive (1), give (1), go (1), grow (1), halve (1), harm (1), have (3), have_to (1), help (1), hurt (1), improve (2), include (1), increase (2), initial (1), institute (1), introduce (1), invest (1), issue (1), laud (1), launch (1), lead (3), leave (1), link (1), maintain (1), make (3), market (1), may (1), meet (2), merit (1), name (5), offer (3), operate (1), pay (2), place (2), pose (1), produce (2), prompt (1), protect (2), pursue (2), put (1), raise (1), reach (3), redeploy (1), reduce (1), remain (2), remove (1), report (3), request (1), require (1), resign (3), resume (1), retire (1), run (1), say (26),

schedule (1), seek (1), sell (1), share (1), show (2), snap_up (1), solve (1), specialize_in (1), stand (1), start (2), step_in (1), store (2), succeed (2), take (1), tell (1), total (4), train (1), trigger (1), turn_around (1), turn_down (1), use (1), vow (1), wallow (1), want (2), watch (1), will (11), work (1), would (1)

WSJ (4): account_for (2), adjust (3), advertise (1), agree (1), announce (1), approve (1), argue (1), ask (1), attempt (1), be (27), begin (2), belong (1), book (1), boost (1), bring (1), burn (1), buy (3), can (5), cap (1), cast (1), change (1), cite (1), clear (1), climb (2), clobber (1), come (1), conform (1), contract (1), contrast (1), could (1), decline (1), declined (1), diversify (1), do (1), drop (1), ease (1), estimate (1), exclude (1), expand (1), expire (1), extend (2), face (1), fail (2), fall (9), file (1), find (3), follow (1), get (1), give (1), grow (2), had_better (1), happen (1), have (3), help (1), hold (3), inch_up (1), indicate (1), insist_on (1), intend (1), invest (2), issue (2), jump (1), jump_in (1), kick_off (1), last (1), launch (2), lead_to (2), leave (1), lend (1), level_off (1), list (2), look (1), mark (1), may (2), might (1), mirror (1), name (1), note (1), offer (1), open (2), operate (1), outpace (1), outstand (1), owe (1), own (1), pay (4), permit (1), pick (1), pick_up (1), plunge (1), predict (2), provide (1), provoke (1), quip (1), raise (1), range (1), reflect (2), remove (1), repay (1), report (1), represent (1), revive (1), rise (5), run (2), run_out (1), satisfy (1), say (23), schedule (1), scramble (1), see (2), seek (1), seek_out (1), seize (1), sell (2), settle (3), show (1), skyrocket (1), slip (1), spend_on (2), stake (1), step_forward (1), stress (1), stretch (1), suggest (2), surge (2), sweep (1), swing (1), take (1), target (1), tend (1), tie (1), trade (7), turn_away (1), turn_up (1), urge (1), watch (1), whipsaw (1), will (4), work_out (1), would (2)