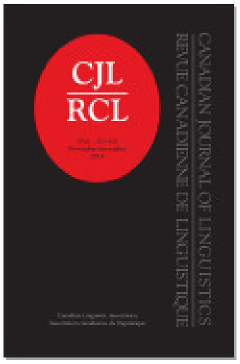
What can adult speech tell us about child language acquisition?

Marjoleine Sloos, Jeroen van de Weijer


CJL
RCL
CANADIAN JOURNAL OF LINGUISTICS
REVUE CANADIENNE DE LINGUISTIQUE

The Canadian Journal of Linguistics / La revue canadienne de linguistique, 60(1), March/mars 2015, pp. 75-83 (Article)

Published by University of Toronto Press

Canadian Journal of Linguistics/Revue canadienne de linguistique 60(1): 75–83, 2015

What can adult speech tell us about child language acquisition?

MARJOLEINE SLOOS
Aarhus University
JEROEN VAN DE WEIJER
Shanghai International Studies University

1. Frequency information in corpora and its use in acquisition studies

This contribution explores a methodological problem in language acquisition studies. Much research in language acquisition has shown that children use statistical learning as a strategy in the acquisition of their native language (Saffran et al. 1996 and many others). Frequency of occurrence is also believed to determine the order of acquisition of phonological structures in the construction of the grammar (Boersma and Levelt 2000, Levelt et al. 2000, van de Weijer and Sloos 2013). How do we obtain the relevant frequency information for acquisition studies?

Ideally, we should take into account children's speech or child-directed speech (CDS), depending on the purposes of the investigation. Investigations into the construction of the lexicon and acquisition of the grammar depend on the input, the perception, and the lexical storage of the child, and therefore, frequency data on CDS seem most desirable. CDS has been investigated to study, for instance, (i) the effect of frequency on the acquisition of verbs and verbal inflection in English (Cameron-Faulkner et al. 2003), (ii) phonological variation in Tyneside English (Foulkes et al. 2005), and (iii) the joint effects of frequency and markedness on language acquisition (Stites et al. 2004). Unfortunately, frequency counts of CDS are not widely available. Although this gap is slowly being filled (for example as a result of the CHILDES database project (MacWhinney 2000)), many languages are still underrepresented, and some data are collected only for specific purposes. This has led researchers to consult other sources, like adult-directed speech or text corpora. This paper addresses the question of whether it is methodologically adequate to use adult-directed speech (ADS) in such language acquisition studies.

We would like to thank two very helpful *CJL* reviewers for their comments on this article. We would also like to thank Joost van de Weijer, and one audience at Shanghai International Studies University, for comments and discussion of an earlier version.

Of course, it very much depends on the topic of investigation whether ADS can be used in L1 acquisition research. As is well known, CDS differs considerably from ADS. For example, nouns are used more often in CDS than in ADS and the reverse holds for verbs (that is, nouns are acquired before verbs (Gentner 1982)); prosodic differences are exaggerated; sentences are shorter (sometimes only isolated words are used); and new words tend to be located in final position (Dominey and Dodane 2004, Ferguson 1964). If we are interested in the lexico-semantic, syntactic, and morphosyntactic aspects of acquisition, a CDS corpus therefore seems indispensable. This is supported by the fact that Goodman et al. (2008) found only weak correlations between the age of acquisition of lexical items and frequency in ADS corpora. However, we would like to argue that in research of the acquisition of *phonological structures* (such as natural classes of sounds, syllable structure, and tone), an ADS corpus may serve as a good substitute if no (adequate) CDS corpora are available. The reason for this hypothesis is that the frequency of occurrence of different phonological structures is not likely to differ very much between different speech styles (specifically, ADS vs. CDS). Consider, for instance, languages with complex onset clusters or nasal vowels, which are usually assumed to be "marked" kinds of structures and less common than single-consonant onsets and oral vowels. If these occur in 20% and 10% of the words in the lexicon, respectively, we expect, roughly, the same relative occurrence for CDS as for ADS. That is, although in CDS, some words may be simplified, we do not expect *consistent* simplification of phonotactics in CDS.

We are not aware of studies that investigated the relation between CDS and ADS for methodological purposes. The comparability of CDS and ADS is, however, addressed in Levelt et al. (2000), who investigated the order of acquisition of syllable structure (CV, CVC, CCVC, VC) in Dutch, assuming a crucial role for frequency in the order of acquisition of the different syllable structures. They first based the frequency counts of the different syllable types on a written database (CELEX; Baayen et al. (1993)) and found that the order of acquisition did not fully match the relative frequency of occurrence of the syllable structures. When the authors compared the order of acquisition with the frequency of the syllable structures in child-directed speech (a single speaker from the van de Weijer corpus (van de Weijer 1999)), a clear correspondence between the order of acquisition and frequency was found. It may thus seem tempting to assume that child-directed speech is mandatory for frequency counts for acquisition studies. But when Levelt et al. (2000) compared the frequency of the syllable structures of CDS with ADS of the same speaker, they found the two corpora showed highly similar patterns in this respect. Thus, for the distribution of four syllable structures in Dutch, child-directed speech and adult-directed speech turned out to be comparable. Our question here is whether we can replicate and generalize this result — which was based on data of a single speaker — to other languages and other phonological structures.

To investigate this, in the remainder of this paper, we compare a CDS corpus with an ADS corpus of French. We first compare lexical overlap of types and tokens then compare the frequency of the distribution of phonological structures. The following section describes the material. Section 3 provides the results and section 4 contains a discussion and the conclusion.

2. Corpus comparison

We used the frequency data of two corpora, a CDS corpus and an ADS corpus. This section provides information about the computation of the frequency (section 2.1) and about the phonological structures that we investigated (section 2.2).

2.1 Lemma and lemma frequency

An important assumption that is central to a usage-based approach such as statistical learning is that the most frequent words are the most relevant for child language acquisition studies (for example Bybee (2007)). In order to investigate the frequency of CDS and ADS, we selected the 450 most frequent words (excluding proper names) in two corpora. For this particular study, we focused on French and compared data of the CHILDES database (a CDS database (MacWhinney 2000)) with the Gougenheim corpus (an ADS corpus of French (Gougenheim et al. 1956)). We extracted frequency information of the French speech data by all adults in their conversations with children, totalling 22,942 words. The CHILDES database provides the frequency of the surface realisations close to the real CDS, including expressions like *oh là là* and *hmm* (which we excluded from the analysis), but not lemma frequency. The Gougenheim corpus contains 8,774 types (7,995 lemmas, including 312,135 tokens) from conversations with 275 speakers. Frequency is provided as lemma frequency, where lemma is defined as a word with all its inflections (but not derivations). Lemma frequency has been shown to be more relevant in word recognition (Jescheniak and Levelt 1994), which may explain why the Gougenheim corpus provides only lemma frequency. For nouns, the lemma includes the singular and the plural like *chose* 'thing' and *choses* 'things'. For adjectives, the lemma includes masculine and feminine like *petit* 'little, small.MASC' and *petite* 'little, small.FEM' as well as their plural counterparts *petits* 'little, small.MASC.PL' and *petites* 'little, small.FEM.PL'. For verbs, the lemma includes all conjugations. However, for strong verbs, like *être* 'to be' and *avoir* 'to have', the different conjugations differ so much from each other that token frequency seems much more relevant since they have different phonological structures. For instance: *être* has a cluster, but the conjugations *suis* [s$^{\text{ɥ}}$i], *es* [ɛ], *est* [ɛ], *sommes* [sɔm], *êtes* [ɛt], *sont* [sɔ̃] do not, so we decided to treat the whole lemma *être* as having no cluster. For the same reason we did not collapse the token frequencies in the CDS corpus into lemma frequency. Pairs like *bon-bonne* (both treated as the lemma *bon*) are treated as one lemma in the Gougenheim corpus. This means that we cannot tell apart the number of nasalized vowels (*bon*) from oral vowels (*bonne*). In those cases, we applied the numbers to the uninflected form. This inconsistency is a shortcoming that is sometimes unavoidable and a consequence of the design of the available corpus.

As a first step in the comparison between the ADS and CDS corpora, we investigated how many words overlapped in the most frequent 100, 300 and 450 words in both corpora. We more or less arbitrarily selected these three sizes of the lexicon to investigate if the correlation between the ADS and CDS corpora was stable, increased, or decreased during the growth of the child's lexicon before he or she starts

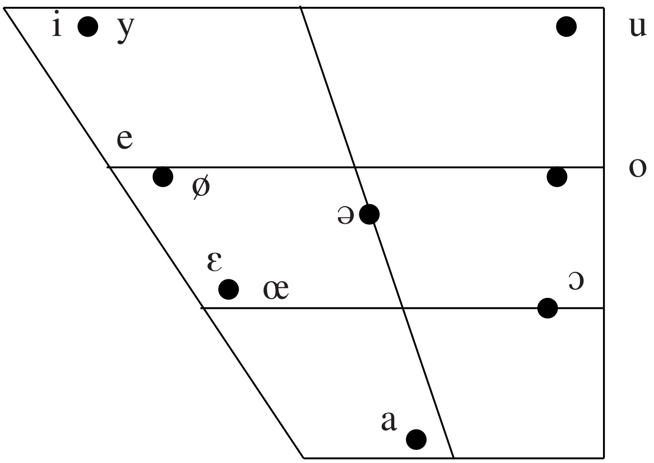


Figure 1: The oral vowels of French (according to Fougeron and Smith 1993)

to speak. At the moment children begin producing speech, they know approximately 500 words (Swingley 2007), which is roughly after one year.

2.2 French phonology

To compare the CDS and ADS corpora, we investigated the frequency of occurrence of seven relevant phonological structures in French (from France), focusing on natural classes and syllable structure (see also van de Weijer and Sloos 2013). The natural classes we considered were nasal vowels, front rounded vowels, and segments with secondary articulation. The vowel system of French is illustrated in Figure 1.

As Figure 1 shows, French has three front rounded vowels /y ø œ/ as in *pure* [pyr] 'pure', *ceux* [sø] 'those', and *oeuvre* [œvr] 'oeuvre'. French also has four nasal vowels /ã õ ɛ̃ œ̃/ as in *sans* [sã] 'without', *son* [sõ] 'his', *vin* [vɛ̃] 'wine', and *brun* [brœ̃] 'brown'. French has the glides [j ɥ w] which can occur as secondary articulation as in *bien* [bʲɛ̃] 'well', *suis* [sᶣi] 'be.1SG', and *avoir* [avʷar] 'to have'. Syllable structure in French allows for onsetless words like the indefinite masculine article *un* [œ̃] and for consonant clusters like in *plume* [plym] 'feather'. Finally, French has codas, where we distinguish coda obstruents *avec* [avɛk] 'with' and coda sonorants *bonne* [bɔn] 'good, well.FEM'. This distinction between the different codas is relevant since in many languages, coda sonorants are allowed while coda obstruents are not (see van de Weijer and Sloos 2013 for more discussion and Fagyal et al. 2006 for a general overview of French phonology and phonetics).

Thus, in the remainder of this article, we investigate the frequency distributions of French front rounded vowels, nasal vowels, secondary articulation, onsetless syllables, consonant clusters, coda obstruents, and coda sonorants, in a CDS and an ADS corpus.

3. Results

In this section, we present results regarding whether the CDS and ADS corpus show lexical overlap (section 3.1), their frequency correlation (section 3.2), and the ranking (that is, the order based on frequency) of the phonological structures and its relation to age of acquisition (section 3.3).

3.1 Lexical overlap

As a first step, we investigated how many words of the ADS corpus also occurred in the CDS corpus. The overlap in types was rather low and decreased as larger parts of the lexicons were considered. For the 100 most frequent words in the ADS corpus, only 64.0% also occurred in the 100 most frequent words of the CDS corpus. For the 300 most frequent words in the ADS corpus, 49.3% also occurred in the 300 most frequent words of CDS corpus. For the 450 most frequent words in the ADS corpus, only 45.3% also occurred in the 450 most frequent words of the CDS corpus. This indicates that for studies that focus or rely on acquisition of lexical items, the ADS corpus cannot be used as a substitute for the CDS corpus. If we compute the overlap of tokens, results get considerably better: 77.8% for the 100 most frequent words, 84.6% for the 300 most frequent words, and 84.3% for the 450 most frequent words. This suggests that the frequency of phonological structures might be comparable in both sets of data, which we will investigate in the next section.

3.2 Comparison between phonological structures in the CDS and ADS corpus

We transcribed the 450 most frequent words in both corpora and marked them for the seven phonological structures we investigated: front rounded vowels, nasal vowels, secondary articulation, onsetless syllables, consonant clusters, coda obstruents, and coda sonorants. Subsequently, we summed the instances of each structure in the two corpora separately. We used the weighted sums: the product of the number of structures that occur in a word and the lexical frequency of that word. For instance, in the Gougenheim corpus, *enfin* [ãfɛ̃] ‘at last, finally’ has a frequency of 1001. The words contain two nasal vowels and one onsetless syllable, so we attributed the value $1001 \times 2 = 2002$ for nasal vowels and $1001 \times 1 = 1001$ for onsetless syllables. Subsequently, we log transformed these frequency counts, since log frequency resembles human perception better (for example, a word that occurs ten times more often than another word is *perceived* as twice as frequent (Shapiro 1969)). Table 1 illustrates the counts for the 100, 300, and 450 most frequent words in both corpora.

Note that, in this order, CDS and ADS for the 450 most frequent words differ only in that the order of secondary articulation and front rounded vowels is reversed (underlined). In the smaller lexicons, coda sonorants are also differently ranked in CDS and ADS. To investigate whether the numbers of structures in the CHILDES and Gougenheim corpora correlate significantly, we conducted a correlation test (in the R statistical environment (R Development Core Team 2009)). We observed a very strong and significant correlation (r = 0.89, $p = 0.003$ for the 100 most frequent words, r = 0.87, $p < 0.005$ for the 300 as well as 450 most frequent words). This

Table 1: The weighted log values of phonological structures in the CDS and ADS corpora

	100 words		300 words		450 words	
Structure	CDS	ADS	CDS	ADS	CDS	ADS
Coda Obstruent	4.29	3.48	4.75	3.98	4.81	4.07
Complex cluster	4.18	3.71	4.72	3.99	4.79	4.07
Secondary articulation	4.71	4.40	4.84	4.48	4.88	4.50
Coda Sonorant	4.78	4.18	4.88	4.29	4.92	4.32
Front rounded vowels	5.07	4.15	5.13	4.27	5.14	4.29
Nasal vowels	5.15	4.39	5.26	4.60	5.29	4.65
Onset	5.45	4.79	5.50	4.84	5.51	4.85

shows that the frequencies of the phonological structures in the CDS and ADS corpora are very similar.

3.3 Ranking concordance on the basis of frequency

For studies that relate the order of acquisition to frequency (like Levelt et al. (2000) and van de Weijer and Sloos (2013)), it is also relevant to compare the *ranking* of the phonological structures based on their frequency, that is higher ranking (higher frequency) is believed to correspond to earlier acquisition. Table 2 provides the acquisition order of the structures under discussion for French (based on Brak (2011) and Rose (2000)).

Table 2: Age of acquisition of the phonological structures of five children

	Nasal vowels	Onset	Coda obstruents	Clusters	Secondary articulation	Coda sonorants	Front rounded vowels
Marilyn	1;11		<1;11	2;03-08		2:04-07	
Clara	1;02	1;05	1;07	1;09	1;07	1;07	1;07
Theo		<1;05	2;03	2;05	2;05	2;03	
Madeleine-1	1;04	1;04	1;06			1;07	1;06
Adrien	3;05	2;01	2;05			2;09	2;03

To investigate the concordance of the ranking, we conducted a Spearman's rank correlation test. We found a strong concordance (S = 16, $\rho = 0.810$, $p = 0.022$ for the 100 most frequent words, S = 16, $\rho = 0.810$, $p = 0.022$ for the 300 most frequent words, and S = 20, $\rho = 0.762$, $p = 0.037$ for the 450 most frequent words). As suggested at the end of section 3.2, this shows that the ranking of the phonological structures based on their frequency in the CDS and ADS corpora are similar. Finally, we compared the ranking of structures based on frequency with the order of acquisition as derived from Table 2. We observed (under Table 1) that the ranking of the structures based on the 450 most frequent words in ADS is the same as in CDS except for secondary articulation and front rounded vowels, which are reversed. The order of acquisition shows that secondary articulation and front rounded vowels are acquired at the same time. So both corpora make the same predictions with respect to the order of acquisition.

Summarizing, although the lexical overlap in types between the corpora is small, the correlation between the frequencies of the phonological structures in the two corpora is very strong. Similarly, the concordance in the ranking of these structures based on their frequency is strong. Finally, both corpora make the same predictions as to the order of acquisition.

4. Discussion and conclusion

In this contribution, we investigated to what extent CDS can be compared to adult speech. The question is relevant for language acquisition studies for which CDS corpora are not readily available. Therefore, we compared frequency information of a child-directed speech corpus with frequency information of an adult-directed speech corpus for French. We selected the 450 most frequent words in these corpora and summed the number of a set of particular phonological structures in these words. Although the lexical overlap of the types in CDS and ADS corpus turned out to be rather small, the overlap in tokens is rather high and so is the overlap of phonological structures. We conclude that the phonological patterns of a language do not differ very much between CDS and ADS: CDS and ADS differ predominantly at the (morpho-)syntactic and lexico-semantic levels. Further, CDS is characterized by hyper-articulation and exaggerated prosody but not by systematic changes in segmental phonological structure. The correlation between the frequencies of the selected phonological structures in the two corpora turned out to be very strong. Similarly, we investigated the concordance of the frequency *rankings* between the two corpora and found that this was also strong and that they make the same predictions as to the order of acquisition. Hence, although the CDS and ADS corpora do not share many words, they are comparable at the phonological level. An interesting question raised by an anonymous reviewer is why the frequency of the phonological structures is higher in all cases in CDS than in ADS. This is probably a result of the fact that the ADS corpus contained lemma frequencies which we could not translate back into token frequencies.

From these results, we infer that ADS corpora can be used as an alternative for CDS corpora in language acquisition studies for particular cases — especially frequency-related studies on phonological acquisition. In other words, if we are interested in the acquisition of certain phonological patterns, the moment of acquisition of these patterns, the order of acquisition of these patterns, or the relative frequency of these patterns, we may use ADS if CDS is not available. As a goal for future research, we advocate more comparative corpus studies (for example, also internet-based ones) to be better able to define their full potential for acquisition studies.

REFERENCES

Baayen, Harald, Richard Piepenbrock, and Hedderik van Rijn. 1993. The CELEX lexical data base on CD-ROM.

Boersma, Paul and Clara Levelt. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of the 30th Child Language Research Forum*, ed. Eve Vivienne Clark, 229–237. Stanford: Center for the Study of Language and Information (CSLI) Publications.

Brak, Jantina. 2011. L'acquisition segmentale et suprasegmentale des enfants français. Master's thesis, Utrecht University.

Bybee, Joan L. 2007. From usage to grammar: The mind's response to repetition. *Language* 82:711–733.

Cameron-Faulkner, Thea, Elena V.M. Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science* 27:843 –873.

Dominey, Peter Ford and Christelle Dodane. 2004. Indeterminacy in language acquisition: The role of child directed speech and joint attention. *Journal of Neurolinguistics* 17:121–145.

Fagyal, Zsuzsanna, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A linguistic introduction*. Cambridge: Cambridge University Press.

Ferguson, Charles Albert. 1964. Baby talk in six languages. *American Anthropologist* 66:103–114.

Fougeron, Cécile and Caroline L. Smith. 1993. Illustrations of the IPA: French. *Journal of the International Phonetic Association* 23:73–76.

Foulkes, Paul, Gerard Docherty, and Dominic Watt. 2005. Phonological variation in child-directed speech. *Language* 81:177–206.

Gentner, Deidre. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In *Language development*, Vol. 2: *Language, thought and culture*, ed. Stan A. Kuczaj, 301–334. Hillsdale, NJ: Erlbaum.

Goodman, Judith C., Philip S. Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language* 35:515–531.

Gougenheim, Georges, René Michea, Paul Rivenc, and Aurélien Sauvegot. 1956. *L'élaboration du français élémentaire*. Paris: Didier.

Jescheniak, Jörg D. and Willem Johannes Maria Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20:824–843.

Levelt, Clara C., Niels Olaf Schiller, and Willem Johannes Maria Levelt. 2000. The acquisition of syllable types. *Language Acquisition* 8:237–64.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rose, Yvan. 2000. Headedness and prosodic licensing in the L1 acquisition of phonology. Doctoral dissertation, McGill University.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274:1926–1928.

Shapiro, Bernard J. 1969. The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior* 8:248–251.

Stites, Jessica, Katherine Demuth, and Cecilia Kirk. 2004. Markedness vs. frequency effects in coda acquisition. In *Proceedings of the 28th Annual Boston University Conference on*

Language Development, ed. Alejna Brugos, Linnea Micciulla, and Christine E. Smith, 565–576. Somerville, MA: Cascadilla.

Swingley, Daniel. 2007. Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology* 43:454–464.

van de Weijer, Joost. 1999. Language input for word discovery. Doctoral dissertation, Katholieke Universiteit Nijmegen.

van de Weijer, Jeroen and Marjoleine Sloos. 2013. Learning markedness constraints: The case of French. In *Linguistics in the Netherlands*, ed. Suzanne Aalberse and Anita Auer, 188–200. Amsterdam: John Benjamins.