Ph.D. Thesis

Numbers on words.
Analyses of the Danish Longitudinal CDI study

Sonja Wehberg

Institute of Language and Communication,

Faculty of Humanities

&

Department of Statistics,

Faculty of Health Sciences

University of Southern Denmark

PREFACE

As indicated on the title page, this Ph.D. thesis is based on a joint project of the Department of Statistics (Biostatistics) at the Faculty of Health Sciences and the Institute of Language and Communication (Center for Child Language) at the Faculty of Humanities. The Ph.D. studentship was financed by the Faculty of Humanities and tied to the large Carlsberg-project "Danske børns sprogtilegnelse".

First and foremost, I would like to thank my "Ph.D. parents" Dorthe Bleses and Werner Vach, who truly showed some amiable parental traits. I would also like to thank all colleagues at the Institute of Language and Communication and the Department of Statistics, especially those who collaborated in writing the articles, proof-read the thesis, came up with title proposals, and/or helped translating parts of it to Danish. Of course, I would also like to thank my family, especially for distracting me effectively when needed.

To add a small personal comment: Danish *is* difficult to learn, not only for children, but also for Germans. In 1927, the German author Kurt Tucholsky wrote "Zum Sprechen eignet sich die dänische Sprache weniger – sie zerschmilzt den Hiesigen auf der Zunge und eilt leichtsilbig dahin, und alles ist ein einziges Wort, und es ist sehr schwer" (in "Eine schöne Dänin"), and nothing much seems to have changed since then. I have to admit, though, that analysing data on language acquisition while self learning the language adds a nice personal flavour to a Ph.D project.

I am very grateful that I got this Ph.D. studentship and with it the unique opportunity to come to Odense: tusind tak for det!

Table of contents

RESUMÉ

Introduction[1]

Child Language Acquisition – offspring of mainly linguistics and (developmental) psychology – is a quite young research area on its own. Its "birth" could be dated 1974, since *The Journal of Child Language (JCL)* was founded then. Six years later, *First Language (FL)* came into existence. Both are regarded as key journals in the field, publishing "articles on all aspects of the scientific study of language behaviour in children, the principles which underlie it, and the theories which may account for it (JCL homepage[2]) and "original research, theoretical articles, review articles and book reviews in all areas of first language acquisition" (FL homepage[3]). A quick browse on the keyword "Danish" resulted in 11 hits in JCL and 9 hits with FL, where almost all of those are related to the research of Plunkett and colleagues (e.g., Plunkett 1984; Plunkett 1986; Plunkett & Strömqvist 1992; see also the Danish corpus in CHILDES supplied by Plunkett[4]): internationally published research on Danish Child Language Acquisition, at least in the past 30 years, is scarce (cf. Bleses et. al, *submitted 2*).

The *Odense Language Acquisition Project* (1998-2001) was founded by an interdisciplinary group of researchers to promote research on first language acquisition in Danish (Basbøll et al. 2002). One part of this project was the adaptation of the American *MacArthur-Bates Communicative Development Inventories (CDI)* (Fenson et al. 1993; Fenson et al. 1994; Fenson et al. 2007), a widely used parental-report instrument in checklist format. This instrument is divided into a form "Words and gestures" (also called *Infant part*), which is targeted at 8-16 months old children, and a subsequent form "Words and sentences" (*Toddler part*) for children between 16 and 30 months of age. The CDI was designed to assess children's early communicative skills in various directions by asking parents, for instance, whether their child

[1] Note that in this resumé, citations from the rest of this Ph.D. thesis, including the papers in chapters 3 to 5, will not explicitly be marked as such.

[2] *http://journals.cambridge.org/action/displayJournal?jid=JCL*

[3] *http://www.sagepub.com/journalsProdDesc.nav?prodId=Journal201667*

[4] *http://childes.psy.cmu.edu*

uses certain gestures (Infant part), understands and/or says certain common words (both parts), or has begun to use more complex constructions (Toddler part). Based on this instrument, the researchers aimed at describing the (average) course of first language acquisition in Danish as well as comparing Danish children's language development across regions, cultures and languages (cf. Andersen et al. 2006).

The data collection strategy in connection with the Danish CDI instrument was two-fold: a large-scale cross-sectional study, comprising a total of 6112 Danish children, was conducted to yield (lexical) norms and allow for analysis of the effect of factors such as *gender* or *parental education* on early language development (Bleses et al., *in press*; Bleses et al., *submitted 1*; Bleses et al., *submitted 2*). This was accompanied by a longitudinal study to be able to truly study the course of development within children. With about 180 participants, the Danish Longitudinal CDI study is a large-scale study as well and presents, to my knowledge, the largest data source among the CDI-based longitudinal studies so far: it is a unique dataset offering almost countless opportunities for analysis.

To assist in analysing the collected data, an interdisciplinary Ph.D. project, resulting in the Ph.D. thesis at hand, was initiated in 2003, and its broadly formulated objective was "to develop and apply statistical methods appropriate to generate and validate hypotheses about language acquisition among Danish children".

The data

The main aim of the cross-sectional data was to establish norms and compare Danish children's average early language development cross-culturally and -linguistically based on measures, which were pre-defined by other published results, that is, basically sum scores. This will not be the focus here: the source of all research presented in the following is data from the Danish Longitudinal CDI study. Based on the Danish CDI instrument, 183 Danish children were followed monthly from 8 to 30 months of age, using the Danish Infant part of the CDI "Ord og gestikulation" between 8 and 15 months of age, and the Danish Toddler part "Ord og sætninger" from month 16 on until study end. The focus of the analyses lies mainly on the vocabulary checklist parts of the CDI, more precisely, on the 410 words (items) contained in both the Infant and the Toddler part. The longitudinal nature of the data allows to trace the emergence of "first words" in a child, and thus, the analysis of acquisition times becomes possible, if one loosely equates acquisition with first reported utterance. The actual definition of the acquisition time of a specific item, that is, the time where an item is first produced (checked off as "said and understood" by the parent), can be carried out on at least two different time scales: the *chronological age scale* as month of first occurrence (e.g., in month 12), or the (individual) *vocabulary size scale* as rank in the acquisition sequence of a child (e.g., as 56th word).

Objectives

As an interdisciplinary project combining both child language acquisition and statistics, the objectives of this thesis were two-fold. The first objective, representing language acquisition, was to gain insight into and document results on Danish children's very first steps into language. Dividing this general objective into several "smaller" ones, we began by describing Danish children's first words and compared them cross-linguistically, then we studied group as well as individual variation: do these first words depend on factors as gender, say, and are there children exhibiting (individual) preferences for some word classes? The last language-related research question concerned the relationship between word pairs: can we identify pairs which are learnt closer together than expected, and do these share some identifiable linguistic features, as being rhyming pairs, for instance? The second class of objectives was of more methodological, statistical nature. Here, we wanted to explore the potentials of the CDI beyond the calculation of sum scores: does analysing single items of the vocabulary checklist, especially with time-to-event methods on the vocabulary size scale, prove fruitful? A further statistical exercise was the development of a measure to formally quantify the extent of closeness for a pair of words.

Results and discussion

First, we were able to re-produce "typical" results on first words with respect to semantic-pragmatic content and sound structure, for example. In light of findings that indicate Danish children typically lag 2-3 months behind their American peers in early vocabulary comprehension and production scores (Bleses et al., *submitted 2*), this was reassuring since we found no support for the hypothesis that Danish children might follow a substantially different "path into the early lexicon" than their American or Italian peers. On the single-word level, however, we identified some striking cross-linguistic differences giving rise to re-consider hypotheses on the character and emergence of first words in general. Staying on the single-word level, we were also able to find differences in average acquisition times (as measured on the vocabulary size scale) relating to groups defined by *gender* and *sibling status*, and we formulated some ad-hoc categories accounting for these differences. Concerning individuality of the lexicon's composition, we found that some children showed time-persistent preferences for some word classes at this early stage. These individual differences between children might be linked to commonly used distinctions such as *referential* versus *expressive* (cf. Shore 1995). Investigating the interrelation of acquisition times of (pairs of) specific CDI items, we obtained results suggesting that words sharing a common semantic-pragmatic feature such as being body parts, names for relatives or (zoo) animals tended to be reported close together. Applying a more refined, relative measure, we also found some interesting patterns. However, these are not as easily linkable to the content of the words as the results for the simpler method. Among others, the hypothesis, that "length of the word" plays a certain role as connecting factor, seems plausible and should be studied further.

Moving on to the second class of objectives, the results of all analyses supported the usefulness of the CDI beyond the usual sum score analyses, despite all its widely discussed limitations (e.g., Pine 1992). Especially, the CDI measure could capture differences on the single word-level as well as individual variation over time. With longitudinal data as ours, the application of methods from time-to-event analysis was possible and promising: as a first illustration, the Cox proportional hazards model on the vocabulary size scale worked well to identify single items which differed in acquisition times between groups; further, we proposed a measure of closeness for a word pair based on the absolute difference of the corresponding acquisition times, relative to the expected difference under (conditional) independence. This measure might also be useful in other settings, where quantifying the extent of *closeness* of two event times is of interest.

Conclusion

Analysing data from the Danish Longitudinal CDI study from two angles, language acquisition and statistics, proved inspiring to both sides. In language acquisition, usually a lot of effort, time, and money goes into data collection. To develop and apply (statistical) methods which can help extract as much information as possible out of the data seems effort-effective, and has shown to bear fruit here: our analyses yielded interesting and meaningful results. Speaking for statistics, statistical science lives off data. The Danish Longitudinal CDI study is in many ways a unique dataset, and to find appropriate ways of addressing research questions within this framework, has not only been exciting in itself, but might, in turn, result in statistical methods applicable in other areas as well. However, looking at the broad topic of the Ph-D. project stated above, I have to conclude that – while it was fun – we have barely scratched the surface yet.

Organisation of this thesis

This thesis is organised in five chapters. Chapter 1 provides an extended summary of the thesis, which sets the results of the subsequent single papers (chapters) into a broader perspective. Chapter 2 begins by introducing the data from the Danish Longitudinal CDI study in detail and highlights then various aspects connected to validity and reliability of the Danish CDI in a longitudinal setting. Next in this chapter, test-retest correlations and predictive values as characteristics of the Danish CDI are presented. Chapter 3 comprises the first paper *Danish children's first words – Analysing longitudinal data based on monthly CDI parental reports* (Paper 1), and chapter 4 includes the companion paper *Girls talk about dolls and boys about cars? Analyses of group and individual variation in Danish children's first words* (Paper 2), together with an addendum containing some technical and theoretical considerations to the analyses described in Paper 2. The third paper *Measuring closeness of two event times* is contained in chapter 5. A Danish resumé constitutes the final section before the Appendix, where sample scans of the Danish CDI forms are provided.

References

Andersen, C., Vach, W., Wehberg, S. & Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Basbøll, H., Bleses, D., Cadierno, T., Jensen, A., Ladegaard, H.J., Madsen, T.O., Millar, S., Sinha, C. & Thomsen P. (2002): The Odense Language Acquisition Project. *Child Language Bulletin 22*, 11-12.

Bleses, D., Vach, W., Wehberg, S., Faber, K. & Madsen, T.O. (in press): *Tidlig kommunikativ udvikling. Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI-forældrerapportundersøgelser af danske normalthørende og hørehæmmede børn* [Early communicative development. An instrument to describe language acquisition of Danish normal hearing and hearing impaired children based on CDI parental-report studies]. University Press of Southern Denmark, Odense.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 1): The Danish Communicative Development Inventories: validity and main developmental trends. Submitted to *Journal of Child Language*.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 2): Building the case for a Danish delay: Early vocabulary development in Danish children in a cross-linguistic CDI-based comparison. Submitted to *Journal of Child Language*.

Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethik, S. & Reilly, J. (1993): *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Singular Publishing Group, San Diego.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D. & Pethik, S. (1994): Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development 59(5)*, Serial no. 242.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J.S. & Bates, E. (2007): *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* Second edition. Paul H. Brookes Publishing Co, Baltimore.

Pine, J. (1992): How referential are "referential" children? Relationships between maternal-report and observational measures of vocabulary composition and usage. *Journal of Child Language 19*, 75-86.

Plunkett, K. (1984): Projekt barnesprog: some preliminary analysis and discussion. *First Language 5*, 219-220.

Plunkett, K. (1986): Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology 27*, 64-73.

Plunkett, K. & Strömqvist, S. (1992): The acquisition of Scandinavian languages. In D. Slobin (ed.): *The crosslinguistic study of language acquisition*, Vol. 3. Lawrence Erlbaum Associates. Chapter 7, 457-556.

Shore, C. (1995): *Individual Differences in Language Development.* Sage Series on Individual Differences and Development. Sage Publications.

Chapter 1

Thesis review

Please note that in this review, citations from the rest of this Ph.D. thesis, including the papers in chapters 3 to 5, will not explicitly be marked.

1.1 Introduction

That one plus one equals two, is something everybody agrees on, aside from some mathematicians perhaps, who might come up with one plus one equalling zero, or something similar. Following Piaget (1952), children at six to seven years of age might have already acquired the notion of *conservation of number*, having passed the necessary sensori-motor period and the pre-operational stage, implying object permanence, symbolic use, and some ability for classification. More recently, children as young as five months of age are found to know that 1+1=2 (Wynn 1992a; cf. also Xu et al. 2005), suggesting "that humans are innately endowed with arithmetical abilities" (Wynn 1992b: 749). Strangely, counting things, possibly a "by-product" of the relentlessly classifying and organising human brain, seems a very natural thing to do, even when real life experience is contradictory in many examples: bringing things together usually involves changing them simultaneously, since they begin to interact in some sense.

Let us look at some examples: some say, adding an apple to a pear gives just two pieces of fruit, but it also increases the probability that both suffer from an early decay, since both apples and pears have a high ethylene sensitivity and ethylene production rate, inducing further maturation[1]: adding an apple to a pear gives two foul pieces of fruit. Adding a new child to the world, if one may put it this way, will most certainly change the child as s/he probably acquires a language, for one thing. However, there is a (slight) chance that the world might change, too. The type of language learnt might influence the child, but the language itself might be changed as well, albeit in most cases almost unnoticeably, since the child might add more weight to beloved words and expressions, help form new structures, or even invent new words.

In this Ph.D.-thesis, we are not working along the lines of a single classical hypothesis-test within a specific theoretical approach. Here, we rather explore what happens if we bring (bio-) statistical methods together with a unique dataset on language acquisition, based on parental reports. I hope to show that the result is not a simple sum, but rather a synthesis: language acquisition profits since some existing hypotheses can be re-considered[2], thus shedding new light on corresponding theories, and some new hypotheses will be generated, giving rise to new research; statistics, on the other hand, profits since, motivated by this intriguing dataset and its linked research questions, new suitable methods have to be developed and applied, which later might prove useful in other research areas as well. In short, I hope to show that here one plus one is more.

1.2 Empirical research in language acquisition

Instead of beginning at the core, that is, describing different theoretical foundations of child language acquisition to gain an overview over the research in this area, we begin at the bottom or, maybe better, the surface of the research: the data. Even if there is no clear-cut starting point in this field, there is no doubt that research on child language acquisition[3] has a long empirical tradition: Berko Gleason, in her introduction to "The Development of Language" (2005: 25) mentions the Greek historian Herodotus reporting on a language acquisition study

[1]cf. *http://www.tis-gdv.de/*

[2]Of course, the character of the given dataset influences the range of hypotheses to be looked at.

[3]Note that the terms "child language", "(first) language acquisition", and "child language acquisition" are used interchangeably here, unlike, e.g., in Ingram (1989: 27).

which the Egyptian king Psammetichus conducted, and gives in Berko Gleason & Thompson (2002: 391) the more recent example of Darwin studying one of his own sons in 1877. In Ingram & Le Normand (1996), data on king Louis XIII's development are analysed, which date back to the beginning of the 17th century. Ingram (1989: 7) himself links the beginning of research on language acquisition to a growing interest in child development at "over one hundred years ago [...], led in many ways by the work of G. Stanley Hall in North America and William Preyer in Europe." (cf. Preyer 1889). He also states that "the active publication of baby biographies can be dated from 1876 with the publication of H. Taine's paper" (page 8; cf. Taine 1877). As a German example, Schaner-Wolles (2001: 224) cites the work of Clara and William Stern (1907) as a "Höhepunkt der psychologisch orientierten Kindersprachforschung jener Zeit"[4].

What all data at all times, which are collected to support research on first language acquisition, have in common, is that they all are measurements, recordings of some sort, of a child's "output" (covering gestures, directions of looks, babbling, or words, for instance). This can be accomplished in different ways: methods span from taking anecdotal notes, keeping a detailed diary over performing experiments to the use of modern technology enabling close sampling strategies for spontaneous speech. The different methods can be categorised along various axes: for example, data collection environment (naturalistic observation versus controlled experiment), length of study period (cross-sectional versus longitudinal), child interpreter (parental-diary data versus transcription of language sample by trained personnel in a standardised manner[5]), sample size (number of children as well as organisation of language sampling, e.g., number and distribution of sampling sessions), degree of standardising of the data (unstructured diary data versus checklist measures, cf. section 1.3 below), and other design parameters.

Theories go hand in hand with data collection formats. Loaning the chronological research categorisation of Ingram (1989: 7-31), child language research begins with the "period of diary studies" which lasted from 1876 until 1926 (Ingram 1989: 7). Most of these longitudinal stud-

[4] Translates to [at this time, the peak of child language research with a psychological orientation].

[5] cf. MacWhinney 2000; Berko Gleason & Thompson 2002 on the communication databases CHILDES or TalkBank

ies were conducted by parents, describing their child's development and focussing on changes of language use, which were of particular interest to them. In this period, a large body of data on child language acquisition was accumulated. Despite their vastness, however, most of this data has been of limited use since the recordings and the quality differed immensely.

Simultaneous to the rise of (developmental) psychology, the "period of large sample studies (1926-1957)" began (Ingram 1989: 11). These studies were mostly driven by a behaviouristic view on language acquisition, where the child was no longer regarded as playing a creative and active part, but rather reacting to stimuli of the environment (e.g., Skinner 1957). In connection to this view on language learning, an increased emphasis was put on systematic research, methodology and measurement. Typically, a large number of children were included in the (mostly cross-sectional) studies, allowing to calculate summary statistics and to establish norms for language development. Even though much effort has been put into the design of these studies, the re-use of their naturalistic data today is complicated by, for example, their use of age-grouped data and their lack of recording equipment – it is difficult if not impossible to extract the original single-subject data to re-analyse them by modern standards.

In 1957, Chomsky's *Syntactic structures* was published and marked the passage to "longitudinal language sampling" (Ingram 1989: 21+23). In these studies, the language development of a small number of children, who usually were not the own offspring of the researchers but carefully selected for study purposes, was studied extensively (e.g., Bloom 1970; Brown 1973). Recording, transcription, and sampling strategies became an issue (cf. Tomasello & Stahl 2004 on effective sampling). Since a main interest lied in establishing "how the child acquires rules of sentence formation" (Ingram 1989: 23), the data should enable going "beyond dating the appearance of two-word utterances to writing rules of their structural properties" (Ingram 1989: 23). Further, "theorists who follow th[is] linguistic approach argue that language is innate in humans" (Bohannon & Bonvillian 2005: 242), implicating that all children share a common language development. Consequently, there was a renewed focus on studying extensively the language development within a single child as opposed to analysing grouped data in large samples. Research concentrated on exploring the "knowledge" of children instead of their "performance" (cf. Chomsky 1965). Therefore, "since 1960, much ink has been spent on issues of experimental design" (Bennett-Kastor 1988: 2), and the area should probably be named one of "longitudinal language sampling or controlled experiments". Curiously, Bennett-Kastor (1988) in her investigation of the practice of child language research (chapter 7) found the following relationship between researchers' affiliation and data collection setting (for the period 1970-1980, roughly): studies based on longitudinal naturalistic language sampling were mainly due to linguists, whereas researchers with affiliations to psychology departments performed experiments (cf. also page 2). Here, one could have expected the opposite, that is, Chomsky's "linguistic approach" (Bohannon & Bonvillian 2005: 241) having a larger impact on linguists, resulting in more experimental studies, and psychologists following their tradition of naturalistic observation (cf. Bennett-Kastor 1988: 2).

Following the categorisation of theoretical approaches to language acquisition in Bohannon & Bonvillian (2005), the large class of "interactionist approaches" constitutes the third main category, besides the "behavioral approaches" and "linguistic approaches". Here, Jean Piaget's cognitive approach and its successors as well as the social interaction approach, which can be traced back to Vygotsky (1962), play a vital role. With respect to data, "the major influencing factor of cognitive or interactionist frameworks upon data was the requirement of contextualization" (Bennett-Kastor 1988: 24). However, the contours of theoretical approaches and disciplines became blurred: "in the current era [1980-1987], the practice of CLR[6] is [...] without the obvious disciplinary alliances of the earlier years" (Bennett-Kastor: 109). But even though "disciplinary chauvinism is by no means evident" (Bennett-Kastor 1988: 110), theoretical foundations continue to interplay with study designs and data collection formats. For example, longitudinal studies on a small number of children prove meaningful when the children are considered to follow a common development. On the other hand, studies on individual variation are only sensible and interesting when working within a framework allowing for true individual variation.

In short, data is intertwined with theory, since one's theoretical approach may influence every step of data handling: collecting, transcribing, coding, measuring, analysing, and interpreting (Bennett-Kastor 1988: 55). And, as stated earlier, all kinds of data analysis are preceded by

[6][Child Language Research]

recording the output of the child and, therefore, a result of interaction of some sort: analysis always requires interpretation (explicit or implicit), either of the child while recording the data or, later, in handling and analysing the recorded data. This is most obvious for diary data where only the result of the communication between child and diary-keeper qualifies to enter the notes. It is less obvious for data based on naturalistic language samples, but nevertheless correct – here, the interpreter is, hopefully, less biased and better trained, and at least the inter-transcriber agreement can be measured (e.g., van Geert & van Dijk 2003) but, still, especially in phonetic transcriptions of very young children's "speech", there is more than one way to do it (cf. Vihman & McCune 1994 on how to identify a word). Even for experiments, though less obvious, the part that communication plays should not be underestimated. For example, different instructions may influence the test results considerably, personal likes and dislikes might have an influence on cooperation and compliance of the children, and even within the highly structured framework/paradigm of preferential listening experiments, for example, the reason for the child to listen longer to one stimulus than another is still to be interpreted correctly: is the longer listening time due to certain acoustic features (of one stimulus), aspects of speech processing, a different degree of familiarity with the stimuli, or something completely different?

So if there is no research on language acquisition, which is free of subjective (theoretical or other) contributions, and if an interpreter is needed, at any rate, to retrieve data, why not make use of the unique parental knowledge? But as opposed to the parental diary studies in former times, do it now in a structured and standardised way, to collect also data on children who are not born to language researchers? The next section will give a short overview of the role of (structured) parental reports in language acquisition.

1.3 Parental reports and the CDI

Research on questionnaire design has shown that it is often recommendable to rely on (outsider) informants to supply the information asked for instead of addressing the target subjects themselves (e.g., Sudman 2004). This holds especially when questions are of delicate nature or the subjects themselves are indisposed to answer (correctly) – which is clearly the case when the aim is to collect information on very young children's behaviour. Here, an obvious choice of informants are the children's parents or primary caregivers, since they are supposed to have the closest contact to and greatest insight into their children (of all potential informants, for example, day-care personnel); this approach seems to work well even with adolescent children, in fact, sometimes better than asking the teenagers themselves (e.g., Pavuluri 2007, where the diagnostic potential of the Mood Disorder Questionnaire for adolescents proved highest in the parental report version as opposed to self-administration).

In child language research, standardised parental reports are a relatively recent form of data collection, at least as a stand-alone collection format (cf. Ingram 1989: 22, citing Braine 1963) with a decided focus on language (for an overview of other measures, e.g., Fenson et al. 1994: 6-9). In 1993, the *MacArthur-Bates Communicative Development Inventories (CDI)* (Fenson et al. 1993; Fenson et al. 1994; Fenson et al. 2007), a self-administered, highly structured parental report, was published. Designed to cover various aspects of early communicative behaviour of very young children, it was the "end-product" of over 20 years of research by Elizabeth Bates and colleagues (e.g., Bates et al. 1975; Bates et al. 1988), where the research included data from observations, diary studies, and from free as well as structured interviews with parents. Some "by-product" parental reports (cf. Fenson et al. 1994: 11-12) are still being used in child language research, for example, the *Early Language Inventory (ELI)* in Bornstein et al. (2004b).

The CDI consists of two separate report forms: the first form "Words and gestures", for which the term *Infant part* is widely used, is intended for children between 8 and 16 months of age, and the second form "Words and sentences", also called *Toddler part*, is designed for children from 16 months of age on until about 30 months of age. These age ranges should serve as guidelines rather than as strict limits, and are adapted to children following a normal communicative development (cf. Fenson et al. 2007). A third form for children between 30 and 37 months of age has been developed as well (cf. Fenson et al. 2007), but has not yet become quite as popular as the first two forms. In addition, several short forms have been developed, which are mainly used for screening purposes (Fenson et al. 2000). In the following, a short overview of the contents of both the Infant and Toddler part of the American CDI is presented. The actual CDI forms can be found as a sample in Fenson et al. (1994, Appendix A).

Infant part "Words and gestures". After some introductory close-ended questions on first signs of understanding, the first section is mostly designed to measure receptive and productive vocabulary, represented both by phrases and single words (items) in an aided-recall or, perhaps better, recognition format (cf. Fenson et al. 1994: 9), where the parents have to check off whether their child "understands" or "understands and says" a specific item. The vocabulary checklist in "Words and gestures" comprises about 400 frequent words and is organised in 19 thematic categories, spanning from "early emerging" categories as *Sound effects and animal sounds* and *Games and routines* to more "advanced" categories such as *Action words* or *Pronouns.* The second section is devoted to the more physical aspect of communication, actions and gestures: it consists of 12 "often-sometimes-not yet" and 51 "yes-no" questions organised in five subsections and ends with the only open-ended question in the Infant part, collecting examples where the child pretends objects to be something else.

Toddler part "Words and sentences". This part begins with an extended vocabulary checklist, covering all 396 items from "Words and gestures" and adding 284 new ones. In contrast to the Infant part, only the productive vocabulary is of interest here, that is, parents are no longer asked to mark an item if it is understood by the child (since this is considered too extensive at this stage of development, cf. Fenson et al. 1993: 5). After five additional close-ended questions on the child's word use (e.g., addressing the child's referring to past and future events), the second section considers development after the one-word stage, that is, sentences and grammar. It begins with questions about the use of word endings and regular as well as irregular nouns and verbs. Commenced by a leading question, an open-ended subsection follows, where parents are asked to provide the three longest sentence examples they heard their child say; the child's MLU could be calculated based on this section. In the last subsection, called complexity, the child's use of constructions is evaluated on the basis of 37 items where parents are asked to choose between two alternative forms, a simpler and a more complex version for each item, the one which best reflects their child's usage.

Since open-ended questions are relatively few, these are rather compact forms – the time it takes to complete either Infant or Toddler part, is estimated to be 20-40 minutes (cf. Fenson et al. 2007: 15), a rather short time considering the gain of information on communicative development of the child. Precisely here lies the strength of the CDI instrument: it resembles a parental diary, comprising much of a diary's in-depth information on single words, for example, and it does so in a manageable amount of time. As a standardised measure, it overcomes the disadvantage of a (parental) diary study, that is, depending on the training, quality and focus of the diary keeper: the CDI provides near-diary data but in a form which allows for comparison. Profiting from the long-term parental knowledge, it avoids a typical short-term (language) testing situation for the child, on which other language assessment measures, e.g., the Reynell Developmental Language Scales (U.S. edition; Reynell & Gruber 1990) rely. It also fills in the void between, on one hand, standardised developmental measures from the psychological field (e.g., the Bayley Scales of Infant Development, Bayley 1993) where language or communicative development of the child is only one domain among others, and labour-intensive language sampling on the other. To summarise, the CDI instrument is a time- and labour-effective standardised measure with specific focus on capturing early language and communication skills, profiting from the unique parental knowledge. Along the lines of a psychological tradition, the quality of the CDI instrument, concentrating on sum scores within the single sections, has been evaluated with the classical concepts of internal consistency, reliability, and validity, and found satisfactory (cf. Fenson et al. 1994; Fenson et al. 2007; see also chapter 2).

So what is it, which only the CDI measure can contribute? The CDI instrument is especially suited to collect data of large samples, and due to its format (and since scanning of report forms is possible), handling and analysing large amounts of data is feasible. Among other things, this allows for a valid analysis of variation in early communicative development, and in particular, variation in the lexical acquisition sequence. Large norming studies have been conducted and corresponding results have been published characterising the early normal communicative development of children (cf. Fenson et al. 1993; Fenson et al. 1994; Fenson et al. 2007 for the American norming study). As a consequence, the CDI can be applied to evaluate a single child's communicative development with respect to these established norms. Also, it has been applied in clinical settings to diagnose and quantify delay (cf. Fenson et al. 2007, chapter 3), and might be useful for screening purposes (although special short forms have been developed for screening, cf. Fenson et al. 2000).

Of course, as all data collection methods, the CDI instrument suffers from some shortcomings. Concentrating for now on the vocabulary checklist, there are at least three limitations of CDI data: first, the definition of a CDI vocabulary checklist has involved choices in selecting CDI items out of all possible early utterances of children, which clearly influences the range (and composition) of early words available for study. This becomes especially relevant if one is interested in the analysis of single items instead of sum scores. Further, the items have been grouped into visible ordered categories. This possibly removes other connotations in filling in the forms. Second, there is no information on the precise use of a word in CDI data, no phonological or contextual information and no information on the assigned meaning, flexibility or frequency of usage. And third, even if parents generally are valid and accurate observers of their children (cf. Fenson et al. 1994; Fenson et al. 2007), it is reported that their recall mechanisms work differently for different word classes, that is, they remembered nouns far better than verbs (Tardif et al. 1999; Pine 1992).

So, in short, the general trade-off between the collection of standardised information and the possibility of (large) sample sizes on the one hand and richness of the data on the other hand holds with the CDI instrument as well. But, anticipating here the results of the papers (cf. chapters 3 - 5), the analysis of single items of the CDI vocabulary checklist proved not only to be possible but also enriching – the CDI measure, despite all its limitation, has a use beyond norms and sum scores; it can capture differences at the single word level and capture individual variation over time. This suggests that its potential, that is, the richness of its data, is not yet fully exploited.

Beside the "original" strength of the CDI, its standardised format is especially suited to allow for adaptations to other languages, sign languages included, and as such this instrument has become very attractive for researchers interested in cross-linguistic comparisons. Up till now, at least 30 adaptations to other languages exist (cf. homepage of the MacArthur-Bates CDI[7]), and many more might follow, since, like a snowball effect, its attractiveness and popularity increases with every new adaptation. There exist guidelines and suggestions for how the adaptations should be done (cf. CDI homepage[8]), and for many adaptations, corresponding

[7] *http://www.sci.sdsu.edu/cdi/adaptations_ol.htm*

[8] *http://www.sci.sdsu.edu/cdi/guidelines_adaptations.htm*

cross-sectional norming studies have been conducted. Concerning data analysis, there seems to have been established a "tradition" to calculate and report sum scores (of single sections of the CDI forms), mostly comprehension and production scores, as a basis for comparison, both within the language and across languages[9]. But the process of standardising adaptations, and calculating and publishing results, is far from being completed yet (cf. Bleses et al., *submitted 2*). A standard for cross-cultural comparisons has not been developed yet; for example, only few researchers take into account that parental reporting might vary systematically across cultures due to a different *social desirability* level (cf. Bornstein et al. 2004b).

1.4 The Danish CDI project

The *Odense Language Acquisition Project* (1998-2001), was founded by an interdisciplinary group of researchers to investigate first language acquisition of Danish children systematically from different angles (Basbøll et al. 2002). One part of this project was the adaptation of the CDI to Danish. This instrument was intended as the basis on which the (average) course of first language acquisition in Denmark could be described. In addition, it was planned to compare Danish children's language development across regions, cultures and languages based on the Danish version of the CDI (cf. Andersen et al. 2006). Sample scans of the Danish CDI, "Ord og gestikulation" (Infant part) and "Ord og sætninger" (Toddler part), are included in the Appendix.

The data collection strategy based on the Danish CDI instrument was cross-sectional as well as longitudinal. A large cross-sectional database was built up, comprising 6112 monolingual Danish children with no reported (chronic or severe) illness (cf. Andersen 2006 for more details on study conduct or sample selection). Based on this cross-sectional study, (at least) three different directions of analysis are pursued. First, norms for the early communicative and lexical development of Danish children have been established (Bleses et al., *in press*; Bleses et al., *submitted 1*; see also CLEX[10]) which make it possible to put subsequent CDI results of

[9] Reflecting this use, the CDI has become more a measure of early *lexical* development than of *communicative* development as a whole.

[10] This web project on "Cross-linguistic Lexical Norms" is currently under development. It aims at becoming a common data platform for CDI data, allowing for direct queries on the datasets. In the beginning, it will

individual children into perspective. Second, due to the large sample size of the study and an extensive accompanying questionnaire, analysing the effect of various factors such as *gender*, *parental education*, and *hours in day-care per week* on early language development in Danish becomes possible; preliminary results are available (cf. Bleses et al. 2003; Andersen 2004). And third, the CDI data allows for cross-linguistic comparisons to investigate whether the alleged delay of Danish children in early communicative and lexical development (e.g., Bleses et al. 2002) exists, and if so, in which respect. Results of a large cross-linguistic comparison of CDI studies can be found in Bleses et al. (*submitted, 2*).

The *Danish Cross-sectional CDI study* was, in fact, preceded by the *Danish Longitudinal CDI study*, conducted between March 2000 and February 2002, where the CDI instrument should be filled in at monthly intervals. The aim was to follow, describe and possibly categorise individual development over time. The CDI instrument (in other adaptations) has been applied longitudinally before, but either only a few times, often in a subsample (cf. Fenson et al. 1993; for a pseudo-longitudinal study: Eriksson & Berglund 1999), or the sample size has been rather small (cf. Dale & Goodman 2005, referring to the San Diego Longitudinal Study with 28 participants). With about 180 participants, the Danish Longitudinal CDI study is a large-scale study; in fact, it presents, to my knowledge, the largest data source among the CDI-based longitudinal studies so far and is, as such, a unique data set offering almost countless opportunities for analysis.

Besides the Danish Cross-sectional and Longitudinal CDI studies, a number of smaller investigations connected to the Danish CDI project have been conducted. For example, reliability and validity of the Danish CDI instrument have been explored in several ways, partly based on spontaneous speech data from the *Odense Twin Corpus* (Basbøll et al. 2002) in relation to CDI forms the twins' parents filled in (Bleses et al., *in press*: chapter 2; Bleses et al., *submitted 2*). Considering the same twin database, amongst other sources, a small input database has been created, comprising input frequencies for all items contained in the vocabulary checklists of the Danish CDI (Rehfeldt & Riegels 2003). Another example is the application of the CDI

include data from the Danish and American CDI norming studies. A preliminary version can be found at *http://www.cdi-clex.org*.

in a study on the effect of maternal fish oil supplementation in lactation on developmental outcome of the breast-fed children (Lauritzen et al. 2005). Furthermore, to study early communicative development of deaf and hearing impaired Danish children, a special version of the Danish CDI has been developed and applied in an (ongoing) study (cf. Andersen & Bleses 2006). And last but not least, a large project to develop a nation-wide language screening tool for 3-year olds involves a short form of the Danish CDI (Bleses et al. 2006).

1.5 The chapters: overview of objectives and results

In this thesis, the focus lies on analysing the Danish Longitudinal CDI study. In chapter 2, we present the data and sample of the Danish Longitudinal CDI study in detail and discuss validity and reliability of the Danish CDI, especially in a longitudinal setting. In the chapters 3 – 5, we explore, amongst other things, the possibilities of analysing single items or words from the vocabulary checklists, whereas "usual" CDI analyses focus on sum scores. With longitudinal data, it becomes feasible to analyse acquisition times of single items, borrowing therefor methods from time-to-event analysis. To describe the precise units of analysis, the time scales involved, and the data structure, which are common basis to all analyses presented in the chapters 3 – 5, two short subsections are included here in this overview before summarising the other chapters. Chapters 3 and 4 comprise the papers on Danish children's first words, both to appear in *First Language*: the first is entitled "Danish children's first words: Analysing longitudinal data based on monthly CDI parental reports" and its companion paper "Girls talk about dolls and boys about cars? Analyses of group and individual variation in Danish children's first words". The third paper "Measuring closeness of two event times" is included in chapter 5.

1.5.1 Chapter 2: the Danish Longitudinal CDI study

Based on the Danish CDI instrument, 183 Danish children were followed monthly from 8-30 months of age, using the Danish Infant part "Ord og gestikulation" of the CDI between 8 and 15 months of age, and the Danish Toddler part "Ord og sætninger" from month 16 on until study end. Gender was distributed almost evenly with 52% girls and 48% boys. 36% of the children were first-born (without siblings at study start) and the remaining 64% had at least one (older) sibling (at study start). Families with a higher social class level (longer education and higher-ranked occupation) were over-represented, as is the case with most voluntary CDI studies (cf. Fenson et al. 2007: 34-38 on the use of the CDI instrument in families with low socioeconomic status). Concerning compliance, the drop-out rate was relatively low, with 134 (73%) of the children having at least 20 out of 23 possible observations.

After a short overview of the longitudinal dataset, the following questions are addressed in chapter 2: 1) are the results on internal consistency based on the longitudinal study similar to the cross-sectional results?, 2) how comparable is the demographic composition of the longitudinal sample to that of the general population?, 3) who drops out of the study and how are missing values distributed (where a missing value arises if a whole CDI form has not been filled in or sent at a specific month)?, 4) do parents employ strategies such as "copy and add" in filling in the questionnaires from month to month?, and 5) are the overall longitudinal results, as measured by some selected sum scores each month, comparable to the cross-sectional results?

To summarise the results, we find an overall large initial as well as pointwise similarity between the Danish Longitudinal and Cross-sectional (norming) CDI studies: the internal consistency is satisfactory, the initial composition of the longitudinal sample is quite comparable to that of the cross-sectional sample, and the distributions of some generally used sum scores agree relatively well (pointwise) for each month. We conclude that the Danish CDI within a longitudinal study design works as well as in a cross-sectional design, even on a monthly basis. However, we see some long-term time effects in the Danish study: for example, the number of missing values per months increases over time, some children stop participating in the study (perhaps related to study results), parental reporting changes more or less from month to month, and (production and complexity) growth curves are steeper than the corresponding normed average towards the end of the study period. Therefore, especially the transfer of results on predictive values or developmental profiles to a general Danish population should be done very cautiously.

Next, we present some results on test-retest correlation and predictive values of the Danish

CDI based on the longitudinal study. The correlation coefficients of common sum scores, as an indicator for reliability or stability, prove to be satisfactorily high and are close to corresponding published American results. Predictive values compare also well to published Swedish results, where available (cf. Eriksson & Berglund 2005). However, relatively low positive predictive values indicate that one early CDI measurement at 16 or 22 months is not a too-well predictor of low or high performance six months later – to reliably predict a child's early lexical development, one would need to consider more than one measurement of the child.

All in all, we have no reason to doubt the validity of the Danish CDI in a longitudinal design, that is, "its measuring what it is supposed to" – even though the "what it is supposed to" could be different for an average Danish child (meaning, for example, that a child who participates in a longitudinal CDI study, could generally have a faster than average communicative development, resulting in higher than average scores). However, concepts of overall validity and reliability apply to the analysis of sum scores within the CDI. Since we are analysing single items of the CDI vocabulary checklists in chapters 3 – 5, validity as "its measuring what it is supposed to" would translate to "parents report all items correctly (at all times)", implying reliability as well. To get a first impression of this "single item validity", we look at the *word continuity index*, the relative number of inversions per word (cf. section 2.2.5). However, to gain more precise information, a specific study would have to be set up, comparing experimental or observational data to parentally reported data at the single word level.

1.5.2 Analysis units

Analyses presented in the submitted papers only involve the vocabulary checklist parts of the Danish CDI, more precisely, these 410 words or items contained in both Infant and Toddler part. All items of the checklist are treated equally, sound effects as *mjav (cat sound)* as well as verbs (e.g., *se (to see)*) or questions as *hvad (what?)*, and no adjustments are considered to account for a possible reporting bias (cf. chapter 2).

The longitudinal nature of the data allows tracing the emergence of (first) words of a child, as opposed to cross-sectional studies, where one usually has to resort to age or total vocabulary size as surrogate measures of time of occurrence of words; with longitudinal data, the definition and analysis of acquisition times becomes possible. Here, acquisition time of a specific item is defined according to the *first time* an item is produced (checked off as "said and understood" by the parent), as opposed to define acquisition time as, say, "the time when an item has been produced three months in a row". In that sense, we consider the cumulative lexicon of a child rather than a point-wise monthly lexicon. This adjusts for instable reporting, for instance, switching from "said" in one month back to "only understood" (cf. chapter 2). Note that the time assignment can be done on at least two different time scales: on the *chronological age scale* as month of first occurrence (e.g., in month 12) or on the (individual) *vocabulary size scale* as rank in the acquisition sequence of a child (e.g., as 56th word).

1.5.3 Data structure

We observe the children monthly at the beginning of each month from 8 up to 30 months of age, however, some children have missing observations at some months. Note that all children are born in the first week in June 1999, so that they turn precisely x months old at the observation time points.

Measuring time by chronological age. Events, that is, the time points of first occurrence of a word, can in our data be observed between 9 and (at most) 30 months, since all children have started at month 8, but some dropped off the study before the official study end at 30 months. Information on words reported at time point 8 is left-censored, since we only know the event occurred before month 9 (and after 0 obviously), and information on words not reported until months 30 (or individual observation period) is right-censored: the word could only have been acquired later, if at all. Note that the information "word A is learnt at time point 10" coincides with "word A is learnt in month 10", if one counts the first month of life as first instead of – what is usually done in calculating age – counting it as zeroth month. The expression "word A is learnt when the child was 9 months old", employs the usual definition of (monthly) age, that is, "after turning 9 months old" or "after having lived full 9 months". Hence, "word A is learnt in month 10" could be translated to the more common "word A is learnt when the child was 9 months old". In the following, we will only use the former expressions such as "word A is learnt in month 10".

Since there are some missing monthly observations for some children, we observe interval-censored data as well, that is, we only know the acquisition time point lies in an interval, for example, in month 10 or 11. Obviously, the monthly data observed have an interval-censored interpretation as well: we observe that acquisition times lie in monthly intervals (between the first and last day of a month), where the same intervals are considered for all children and all words.

Measuring time by lexicon size. The lexicon size is defined as number of words reported as present in the CDI vocabulary. For each child, we observe if the event of interest, the first occurrence of a specific word, occurred between the individual lexicon sizes at the beginning and end of a month. As an example, a child produces the word *dukke (doll)* for the first time in month 12; she has already 48 words in her lexicon before month 12, and learns 33 words in this month; the exact time of acquisition on the vocabulary size scale for *dukke (doll)* lies between the 49th and the 81st word. As opposed to chronological age, the intervals differ here between children. Note that in Paper 2, we define the acquisition time on the vocabulary size scale as the first possible time point in a month, that is, 49th for *dukke* (cf. also Addendum to Paper 2).

1.5.4 Paper 1: objectives and results

Here, we focus on the very early phase of lexical development, that is, the period before a child's vocabulary comprises more than one hundred words, dividing this phase into the vocabulary stages *first-1 word*, the *first-5*, *-10*, *-25*, *-50*, and the *first-100 words* of the child. Having monthly data, we can only determine, in which month the xth word was produced, but we cannot identify the xth word itself. Therefore, we refer to all words produced prior to or in the month, in which the xth word was produced, as the first-x words. In the example for *dukke (doll)* from the section above, all 81 words learnt before month 13 qualify as *first-50 words* since the 50th word occurred in month 12.

Our main objective is to study and describe the first words produced by Danish children: we analyse some important aspects of early words – or rather of *early word targets*, that is, the adult models the children attempt to use (Vihmann 1996), namely their 1) sound structure, 2) semantic-pragmatic content, and 3) form based on adult use (e.g., *Common nouns* or *Descriptive words*). Regarding form, we follow the approach of Bates et al. (1994) and Caselli et al. (1995) when defining the categorisation of early words. In addition, we highlight similarities and differences in first words across specific languages and thereby gain insight into particularities of the Danish language and culture. Also, we look into whether there are certain characteristics for words to become first words. For comparison, we use results published on American English and Italian first words, based on cross-sectional CDI data (Caselli et al. 1995).

With respect to the sound structure, we analyse the syllabic structure of adult target words and find mostly monosyllabic words, where the few polysyllabic items (e.g., *mormor (maternal grandmother)* and *banan (banana)*) are characterised by reduplication. The bilabials [m] and [b] account for almost half of the initial consonants of the (very) first words in Danish, fitting well to results reported on American CDI data where "[w]ords beginning with b [...] make up 24% of the first 100 words" (Dale & Goodman 2000: 73-74).

With respect to content, the findings on Danish first words are not surprising as well: the children know names for mother and father, affirmations (*ja (yes)*) and prohibitions (*no*), they use words linked to social interaction contexts such as greeting (*hej (hi)*) and playing (*borte tit (hiding game / peek-a-boo)*), objects (presumably) close to a child's world (*bil (car)* and *bog (book)*) and they talk a lot – using sound effects as well as common nouns – about cats, dogs and the like. However, despite the overall "typicality" of Danish first words, we also find some striking differences in comparing the acquisition ranks of single words between Danish, American English, and Italian, for example, in *bil-car-automobile*, *tak-thank-you-grazie*, *mad-food / mealtime-pappa*, and – most prominent – *mor-mommy-mamma* and *far-daddy-papa*, which Danish children acquire later than their American or Italian peers.

Also with respect to early vocabulary composition, we find "typical" results. Following the approach of Bates et al. (1994) and Caselli et al. (1995) to categorise CDI items based on the original CDI checklist's categorisation into, for instance, *Sound effects*, *Common nouns*, or *Action words*, we find that the very first words of Danish children are mainly *Sound effects*, which account on average for over 50% of the vocabulary at the *first-1 word* stage, followed by *Games and routines* with little more than 25%, and *People* and *Common nouns* more or less equally accounting for the rest. Following this early phase, we see a transition over time towards *Common nouns* at the expense of *Sound effects* and *Games and routines* which gradually lose rank and prominence over time. *Action words*, *Descriptive words*, and *Function words* only begin to pick up noticeably in frequency after the *first-50 words*, implying that at the *first-100 words Common nouns* still dominate Danish children's vocabularies with an average contribution of more than 50%. Comparing the Danish results to American English and Italian results, respectively, some initial differences show up, but they vanish rapidly with increasing lexicon size: the results for all three languages show a strikingly similar development in spite of substantially different languages.

Exploiting the longitudinal nature, we look at the course of first words' ranks over time (from *first-1* to *first-100 words*) and establish three different "ranking patterns" of Danish first words: 1) words with stable rank over the *first-words* stages are used (very early on) at any rate by almost any child; their adult models are mostly easily pronounced social words or frequently used sound effects; 2) words with decreasing rank emerge "early or never" and are games, routines, or more infrequently used sound effects; and finally 3): words with increasing rank can be roughly categorised as simple and important object and person names. The CDI seems here to capture the transition from a prelexical stage in the very beginning to the early labeling phase (cf. Vihman 1996).

1.5.5 Paper 2: objectives and results

As in paper 1, we focus again on Danish children's *first-100 words* (produced). Here, we analyse different aspects of variation in the data. The first research question is whether there are differences in the very early lexicon of children which can be attributed to pre-defined groups due to *gender* or *birth order*. Just asking "are there any words which girls acquire faster (at a younger age) than boys?" would probably result in the answer "almost all words" since girls are reported to be generally (a bit) faster than boys in lexical acquisition (cf. Fenson et al. 1994; Maital et al. 2000; Bornstein et al. 2004a; Szagun et al. 2006). Instead, we analyse the influence of *gender* and *sibling status* (only child versus at least one sibling) on the acquisition time as measured by vocabulary size of single first words. As such, we look at the individual acquisition sequences of children and phrase the research question as "is it possible to identify first words, which boys, for instance, tend to have earlier (as 10th word, say) in their lexicon than girls (who have it first as 50th word, say)?". For this, statistical methods for analysing time-to-event data are used. Additionally, we investigate whether there is an influence of *gender* and *sibling status* on the composition of the vocabulary, that is, on the relative frequency of "word classes" such as *Sound effects* or *Common nouns* as defined in chapter 3, see above.

For each single word and each child characteristic, we fit a Cox proportional hazard model (e.g., Kalbfleisch & Prentice 2002). The acquisition time for a given child is defined as the first word in the corresponding month (cf. section 1.5.2 above). Take as an example the following: a child says *dukke (doll)* for the first time in month 20, among 12 other words appearing for the first time at this time point; before month 20, the child has already produced 51 words: the acquisition time for *dukke (doll)* (and the other 12 words) is set to 52. For children who did not produce the word within their *first-100 words*, the acquisition time is regarded as censored and set to 101 (or to the child's final lexicon size plus 1 if less than 101). For a detailed discussion of the implications of this approach, see Addendum to Paper 2.

We rank the words after their p-values, referring to a test of the null hypothesis that the hazard ratio is equal to 1, instead of, for example, their estimated hazard ratios, since we are interested in identifying words where we are sure that there is a difference and want to avoid including words, where a large estimated hazard ratio is based on (extreme) results of very few children. To assess the overall significance of our findings we have to take the number of analyses performed into account. For *gender*, we analyse all 306 words produced by at least 10 children and find 45 words observed with a p-value below 5% versus 15 expected (as 306x0.05=15.3). For *sibling status*, we include *gender* as an additional factor in the analysis and analyse all 262 words which were produced by at least 20 children within their *first-100 words*; at the 5% level, we observe 30 words vs. 13 expected, hence we have clear evidence for an influence of these factors on the acquisition time of some words.

With respect to semantic-pragmatic content, we sort and structure the identified words by a post-hoc categorisation: girls have, for instance, words focusing on social relations, personality, and words for "objects to be cared for" earlier in their productive repertoire than boys, whereas boys are earlier in naming, for example, loud, moving objects, objects they can act on and certain food-related items. Examples for word groups which first-borns use earlier than later-borns are names for caregivers other than their parents, sound effects such as *vov (dog sound)*, as well as labels for boy and girl. Later-borns, on the other hand, know for instance names for their siblings and for objects and activities usually related to older children (e.g., *slå (to hit)*) earlier than first-borns.

Despite differences on some single words, we find nothing to support the hypothesis of structurally different vocabularies, when we investigate the effect of *gender* and *sibling status* on the composition of the lexicon based on adult-defined formal categories: while children might differ with respect to *what they talk about*, we find at this very early level no difference in *how they talk*, according to their *gender* or *sibling status*.

In a second approach to describe and explain variation, we try to identify groups of children sharing a common pattern of development, thus addressing the question of individual style. We copy the approach of Lieven et al. (1992) and compare the vocabulary composition in the *first-50 words* of a child to that in the *second-50 words*, using our longitudinal CDI data to look for time-persistent preferences for certain word classes within children. These may, if they exist, reflect "individual paths into language" and may also be interpretable in terms of already established distinctions such as referential-expressive, analytic versus holistic, or a preference for frozen phrases versus common nouns (cf. Nelson 1973; Bates et al. 1988; Lieven et al. 1992; Shore 1995).

To do this in practice, we correlate the fraction of words from a certain category (e.g., *Common nouns*) between the first and second 50 words of a child, where the *first-50 words* are defined as the first part, and all new words up to the *first-100 words* as the second part. However, since the CDI measure limits the total range of words as well as that of each category, the fraction of words within the second part may suffer from a ceiling effect, and consequently, there will be a tendency towards negative correlations. We correct for this ceiling effect by estimating for each child and each category the expected fraction in the second part given the fraction in the first part, and base the assessment of a persistent preference for a category over time on the difference between observed and expected fraction instead of just the observed fraction.

We find highly significant positive correlations around 0.3 for *Common nouns*, the combined group *Action, Descriptive, and Function words* as well as for *Function words* alone, and correlations ranging from 0.14 – 0.19 for the other categories, except for *Sound effects*. The correlation for *Sound effects* was -0.03, but changed to +0.11, when we removed one outlier. This result is still a small correlation, but might in part be due to the limitation of our correction approach for very small categories. Overall, these findings indicate that there are moderate time-persistent preferences for certain word classes within the *first-100 words* of children.

1.5.6 Paper 3: objectives and results

The basic purpose of this paper is to develop (statistical) methods which would allow looking for those linguistic properties of a word, which may affect the time point of acquisition into the productive repertoire. As a first step, we want to identify those word pairs from the Infant CDI vocabulary checklist which have a tendency to be learnt close together and, second, we want to quantify the extent of this closeness. Then, it becomes possible to take a closer look at those word pairs with a large closeness value, hoping to identify linguistic word-related features – common to several word pairs – which may explain their closeness. For example, children might learn words closer together than expected if they are connected by a semantic category as "zoo", or a sound-related link as between *is* and *gris* or *faster* and *pasta* (rhymes), or perhaps a formal linguistic-grammatical category (verbs, question words) or others.

A dependency measure such as, for example, the correlation coefficient, does not do the task, since we would obtain a perfect correlation if word A always takes, say, twice as long to learn as word B, but these words are by no means learnt close together: we are interested in *agreement* rather than *association*. Following the approach in Altman & Bland (1983) and Bland & Altman (1986), we propose to measure closeness based on the (absolute) difference between the two acquisition or event times T_1 and T_2.

However, it is well known, that one cannot estimate "the upper tail" of the joint distribution of T_1 and T_2 from censored observations in a non-parametric way. Consequently, we cannot expect to be in general able to estimate the distribution of the absolute difference. However, we can approach the problem by considering a conditional distribution instead, where we condition on that the acquisition time of at least one of the two words is actually observed within a relevant subrange.

This allows us to inspect closeness of the two event times based on absolute measures, such as the probability that the two events occur in the same month. However, since small values of the difference may happen by chance, and the amount of this chance closeness depends on the marginal distributions of the event times, it is hard to judge closeness based on an absolute measure alone. Therefore, we relate the absolute measure to the closeness measure under chance conditions, that is, independence of T_1 and T_2. In addition, we take into account the situation where it might be sensible to consider closeness relative to conditional independence of the two event times given another variable instead of (unconditioned) independence of T_1 and T_2. For example, girls are in general considered to be faster than boys in acquiring words (cf. Fenson et al. 2007). For a *gender*-sensitive pair of words, we would not want to compare closeness relative to (unconditional) independence, ignoring *gender*, because this would artificially enlarge the difference under chance condition, and consequently result in a smaller than justified closeness.

After introducing methods to measure and estimate closeness of two event times based on the absolute difference, we present several possible concrete approaches and apply them in our data. For the simple absolute measures, we find that word pairs appearing close share some interesting features: many pairs consist of body parts such as *arm*, *leg*, or *nose*, and other pairs show an opposing relation, e.g., *no-yes*, *boy-girl*, *grandfather-grandmother*, or *red-blue*. To gain a first impression on the relation between more than two words at one time, we perform a hierarchical cluster analysis and find, for example, 20 clusters of size ≥ 3 with an average pairwise value of the measure ≥ 0.25: the largest cluster consists of 6 words (*hand, leg, bed, arm, head, finger*) and the next largest of 5 words (*hair, eye, nose, mouth, ear*). All other clusters contain 4 (e.g., *tiger, giraffe, elephant, lion*) or 3 words (e.g., *to fall, dirty, to* *run*). Considering th relative measure based on the ratio leads to a more drastic re-assessment compared to the absolute measures – the content of the words (such as being a body part) does not seem to be the most obvious common factor any longer. Again, we perform a hierarchical cluster analysis and find, for example, 4 clusters of size ≥ 3 with an average pairwise value of the relative measure ≥ 3: the largest cluster consists of 10 words (*to watch, (small) box, playpen, rocking chair, nice, pleasant/nice, person/human being, shop/business, sledge, to stop*) and the two next largest of 4 words: *to sit, to show, to hurry, behind* and *deer/stag, pony, turkey, puppy.* The smallest cluster here is of size 3: *to dry, to pull, to take.*

The concept of closeness, as we approach it here, seems to be in principle working well, yielding sensible and interesting results, at least for the simple measures. However, the concrete implementation, especially for the relative measures or when taking covariates into account, certainly needs further work and refinement – this presents an exciting task for the future.

In the following we present a short synthesis and general discussion of the Ph.D. thesis and point to future research topics in this field.

1.6 Synthesis and discussion

As an interdisciplinary project combining both child language acquisition and statistics, the objectives of this thesis also covered both sides, language acquisition and statistical methodology. With respect to language acquisition, we wanted, in general, to gain insight into and document results on Danish children's first steps into language. To do this, we described Danish children's first words and compared them cross-linguistically. In addition, we studied group as well as individual variation: do the first words depend on factors as *gender*, say, and are there children exhibiting individual preferences for some word classes? As a third objective within language acquisition, we addressed the question whether word pairs can be identified which are learnt closer together than expected, and whether these word pairs share some identifiable linguistic features, such as rhyming, for instance. The second class of objectives was of more methodological, statistical nature. Here, we wanted to explore the potentials of the CDI beyond the calculation of sum scores: does the analysis of single items of the vocabulary checklist prove fruitful, in particular with time-to-event methods on the *vocabulary size* time scale? A further statistical exercise was the development of a measure to formally quantify the extent of closeness for a pair of words.

First, we were able to re-produce "typical" results on first words as seen in many other studies (e.g., Nelson 1973; Bates et al. 1994; Dromi 1999; Dale & Goodman 2005). In light of previous findings indicating that Danish children typically lag 2-3 months behind their American peers in early vocabulary comprehension and production scores (Bleses et al., *submitted 2*), this "typicality" was reassuring since we found no support for the hypothesis that Danish children might follow a substantially different "path into the early lexicon" than their American or Italian peers.

On the single-word level, however, we identified some striking cross-linguistic differences. To explain these, cultural differences accompanied by input frequencies and/or linguistic properties of the words seem plausible, confirming that culture and language matter to first words. Let us take a closer look at the most prominent of theses differences, the names for mother (*mor*) and father (*far*). Assuming that cultural differences play no role here, the late acquisition in Danish might be due to the more difficult sound structure in Danish: the Danish terms for parents have no easy consonant reduplication matching common babbling schemes as it is seen in many other languages (cf. Jakobson 1962; Levin 2005). An alternative hypothesis is that instead of Danish children being late, American and Italian children are rated as more advanced than they actually are. To exemplify, a child utters repeatedly [ma] and her American mother recognises this as *mommy* (and not as a highly implausible *mammal*), whereas her Danish mother recognises it as meaning *mad (food/mealtime)* stimulating her child by her reaction to re-use it in meal-related situations. Instead, the child producing [ma] while sitting in the bath goes unrecognised and gets no stimulation. The parents as the usual first communication partners play the important part as (mis-)interpreters; they react with great enthusiasm to potential first words resulting in more and improved attempts of the child. This view is far from being new, other researchers have been focussing on this "dialogic process" (Vihmann 1996: 135) before (e.g., Jespersen 1922/1964; Dore 1983; Veneziano 1981), but here we were able to add new (data) substance to it. Our findings support the hypothesis itself as well as the CDI instrument as method of data collection.

Staying on the single-word level, we were also able to find differences in acquisition times (measured on the vocabulary size scale), which could be attributed to groups defined by *gender* and *sibling status*. For the words showing group-related differences in average acquisition times, we formulated some ad-hoc content-related categories, which might, in the future, give rise to new research exploring their usefulness. Besides generating new hypotheses based on these post-hoc categories, we re-considered some "old" ones since the factors *gender* and *sibling status* themselves have been found to be of (various) impact on language acquisition before, especially so on rate of lexical acquisition (cf. Fenson et al. 1994; Bornstein et al. 2004a, 2004c). Here, we re-viewed the influence of these factors from a different angle and concluded that, even though the effect of speed alone had been cancelled out, the socialisation-related factors *gender* and *sibling status* still had an effect on the first words children produce. We did not find an effect of the same factors on composition of the early lexicon, indicating either that differences in language style – if existent at this very early stage – should be investigated employing a different categorisation approach, or that the effect of these factors is truly limited to semantic-pragmatic content. Hypotheses about possible causes for "why children talk earlier or later about different things" remain to be investigated.

Concerning individuality of the lexicon's composition, we found individual time-persistent preferences for certain word categories even at this early stage in language development, demonstrating the usefulness of the CDI also in this respect. Comparing our results to those published in Lieven et al. (1992), despite the limitations of such a comparison, we found surprisingly similar results on "comparable" categories, such as *Common nouns* and *Interactive words*. This leads to suggest that the preference for *Frozen phrases* (in Lieven et al. (1992)'s framework) could be substituted with a preference for *Action, Descriptive, and Function words* or *Function words* alone in the CDI setting, bridging the two formulations. This, of course, remains to be studied further. It also remains to be addressed whether, and if so how, the variation in vocabulary composition might be used to identify and characterise disjunct groups of children with specific lexical acquisition profiles. Then, one might be able to study, how this variation is linkable to terms such as referential-expressive, analytic versus holistic, or other distinctions (cf. Nelson 1973; Bates et al. 1988; Lieven et al. 1992; Shore 1995).

Investigating the interrelation of acquisition times of (pairs of) specific CDI items, we obtained results suggesting that words sharing a common semantic-pragmatic feature – such as being body parts, names for relatives or (zoo) animals – tended to be reported close together. Applying a more refined, relative measure, which sets the "simple" closeness in relation to the expected closeness under chance conditions, we were also able to detect interesting patterns among the word pairs, which are learnt close together. However, these are not as easily to connect to a common linguistic feature as the results for the simpler method. Here, the analyses based on the proposed closeness measures need to be expanded further to evaluate the role of some linguistic hypotheses: for example, does the factor "length of the word" influence closeness of word pairs?

Considering the second class of objectives, the results of all analyses support the usefulness of the CDI, despite all its widely discussed limitations (e.g., Pine 1992), beyond the usual sum score analyses. Especially, the CDI measure could capture differences on the single word-level as well as individual variation over time. With longitudinal data as ours, we have the unique opportunity to study development within individual children. Consequently, the application of methods from time-to-event analysis has been possible (and promising): as a first illustration, the Cox proportional hazards model on the vocabulary size scale worked well to identify single items which differed in acquisition times between groups.

Analysing data from the Danish Longitudinal CDI study from two angles, language acquisition and statistics, proved inspiring to both sides. In language acquisition – and not only there – usually a lot of effort, time, and money goes into collecting data, and indeed has done so, since there is a lot of ready (CDI) data available at this point in time. To develop and apply statistical methods which can help extract as much information as possible out of the data seems effort-effective, and has here shown to bear fruit, since our analyses yielded interesting and meaningful results: research on child language acquisition profited in both content-related as well as methodological respects.

In turn, statistics might profit since we defined a measure of closeness for a word pair based on the absolute difference of the corresponding acquisition times, relative to the expected difference under (conditional) independence. The concept of closeness might also be applicable and useful in other settings, where quantifying the extent of closeness of two event times is of interest.

From a broader point of view, statistical science gives most meaning in connection with data. The Danish Longitudinal CDI study is in many ways a unique and inspiring data set, and to find appropriate ways of addressing further research questions within this framework, will not only be exciting in itself, but might, in turn, result in more refined statistical methods applicable in other areas as well. So right here is one of the rare occasions where I go from being a mathematician, to whom counting is dear, to being a follower of interactionist theory, hoping that adding the two fields together yields something more to both sides.

1.7 Future research

We explored the data from the Danish Longitudinal CDI study in various ways and came up with interesting results in almost any direction. However, I have to realise that we have barely started yet: my personal list of possible future analyses still exceeds the list of analyses done, by far. In the following, I will shortly sketch a few of these possibilities.

First, on the most concrete level, the work presented in chapter 5 is far from being exhaustive; it could be expanded in various directions. To name four of those, first, the problems we encountered in the concrete implementation need to be addressed. Second, a more extensive cluster analysis based on the different closeness measures between single word pairs could be used to precisely determine patterns between larger groups of words acquired close together by sub-groups of children. Third, the closeness measure could be adapted to also cover the situation, where word pairs are learnt with a constant distance (of 5 months, say), and last, data from other sources could be analysed to explore whether the closeness measure can be of use beyond the application within the CDI.

Second, one could try to do essentially the same based on the cross-sectional CDI data instead, namely, looking for *unknown* linguistic properties of a word, which might determine the time point of acquisition into the productive repertoire. This could be done along the lines of Vach & Bleses (*in prep.*) who studied the association among "acquired yes/no" for pairs of verbs, adjusted for the general proficiency of a child. The results could then be used to put the longitudinal findings in perspective, thus complementing and completing the picture. Or instead, one could attempt to analyse the effect of some *known* linguistic properties, such as the number of syllables or sound structure, on the acquisition time, possibly adjusted for input frequency. Again, this could be done using both the longitudinal and the cross-sectional dataset. Note that all the analyses mentioned so far still focus on the single items of the vocabulary checklist of the CDI.

Third, going from single items to the overall development, the (lexical) acquisition process as a whole could be studied thoroughly, using the longitudinal dataset to identify common patterns or address the question of speed of lexical acquisition and vocabulary spurt (cf. Vach et al., *in prep.*). Here, it would be especially interesting to see whether there exists a general order of words in which children learn the CDI vocabulary – where the order might just be disturbed by some random noise, or whether there exist distinct patterns shared by several children, reflecting true individual variation between children.

And last, besides analysing the vocabulary checklist with respect to perception, other sections of the CDI might indeed be worthwhile to analyse, on their own as well as in connection with the vocabulary checklist. One example of this is the analysis of the single items in the complexity section: to examine whether some grammatical features are more closely tied to the lexicon than others, their acquisition times could be studied, depending on the lexical proficiency of the children. Another example is the comparison of gestural development and lexical acquisition, based on the unique longitudinal dataset, studying whether some gestures are necessary pre-conditions for building up a vocabulary. Furthermore, since information from a small additional questionnaire is available for the Danish Longitudinal CDI study, it would, in principle, also be feasible to investigate whether changes in the development are linked to other outside factors, occurring some time before (e.g., start in a day-care institution).

References

Altman, D.G. & Bland, J.M. (1983): Measurement in medicine: the analysis of method comparison studies. *The Statistician 32*, 307-317.

Andersen, C. (2004): Små børns sprogrejse – En empirisk analyse af danske børns leksikalske tilegnelse i alderen 8-15 måneder [Small children's language journey – an empirical analysis of Danish children's lexical acquisition between 8 and 15 months of age]. Master of Arts thesis. Center of Nordic studies. University of Southern Denmark, Odense.

Andersen, C., Vach, W., Wehberg, S. & Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Andersen, C. & Bleses, D. (2006): Conduct of the Danish CDI-DKDH Studies of Children with Hearing Impairment. *Working Papers, Center for Child Language e-prints*, no. 2. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Basbøll, H., Bleses, D., Cadierno, T., Jensen, A., Ladegaard, H.J., Madsen, T.O., Millar, S., Sinha, C. & Thomsen P. (2002): The Odense Language Acquisition Project. *Child Language Bulletin 22*, 11-12.

Bates, E., Camaioni, L. & Volterra, V. (1975): The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly 21*, 205-226.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J.S., Reilly, J. & Hartung, J. (1994): Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language 21*, 85-124.

Bates, E., Bretherton, I. & Snyder, L. (1988): *From First words to Grammar: Individual differences and dissociable mechanisms.* Cambridge University Press.

Bayley, N. (1993): *Bayley Scales of Infant Development.* Second edition. Harcourt Assessment, San Antonio.

Bennett-Kastor, T. (1988): *Analyzing children's language.* Basil Blackwell.

Berko Gleason, J. & Thompson, B. (2002): Out of the baby book and into the computer: child language research comes of age. *Contemporary Psychology: APA Review of Books 47*, 391-394.

Berko Gleason, J. (2005): The development of language: an overview and a preview. In J. Berko Gleason (ed.): *The Development of Language.* Allyn & Bacon. Chapter 1, 1-38.

Bland, J.M. & Altman, D.G. (1986): Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet*, i, 307-310.

Bleses, D., Basbøll, H., Madsen, T.O. & Thomsen, P. (2002): Kløjs danske børn i sproget? Fokus på danske børns tilegnelse af leksikon [Do Danish children choke on their language? Focusing on Danish children's lexical acquisition]. Studier i Nordisk, Selskab for Nordisk Filologis Årsberetning 2000-2001. 69-81.

Bleses, D., Thomsen, P., Basbøll, H., Madsen, T.O., Vach, W., Wehberg, S. & Nielsen, C. (2003): Midtvejsrapport: Tværsnitsundersøgelse af danske børns ordforråd i alderen 8 til 15 måneder [Intermediate report: A cross-sectional study of Danish children's lexicon between 8 and 15 months of age]. Center for Language Acquisition, University of Southern Denmark, Odense. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Bleses, D., Vach, W, Worm, T. & Møller, M. (2006): Sprogscreening af 3-årige børn [Language screening of 3-year old children]. *Tale-Høre-Nyt*, no. 4.

Bleses, D., Vach, W., Wehberg, S., Faber, K. & Madsen, T.O. (in press): *Tidlig kommunikativ udvikling. Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI-forældrerapportundersøgelser af danske normalthørende og hørehæmmede børn* [Early communicative development. An instrument to describe language acquisition of Danish normal hearing and hearing impaired children based on CDI parental-report studies]. University Press of Southern Denmark, Odense.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 1): The Danish Communicative Development Inventories: validity and main developmental trends. Submitted to *Journal of Child Language*.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 2): Building the case for a Danish delay: Early vocabulary development in Danish children in a cross-linguistic CDI-based comparison. Submitted to *Journal of Child Language*.

Bloom, L. (1970): *Language development: form and function in emerging grammars*. MIT Press.

Bohannon III, J. & Bonvillian, J. (2005): Theoretical approaches to language acquisition. In J. Berko Gleason (ed.): *The Development of Language*. Allyn & Bacon. Chapter 7, 230-291.

Bornstein, M., Hahn, C-S. & Haynes, O.M. (2004a): Specific and general language performance across early childhood: stability and gender considerations. *First Language 24*, 267-304.

Bornstein, M. & Cote, L. with Maital S., Painter, K., Park, S., Pascual, L., Pécheux, M-G., Ruel, J., Venuti, P. & Vyt, A. (2004b): Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development 75*, 1115-1139.

Bornstein, M., Leach, D. & Haynes, O.M. (2004c): Vocabulary competence in first- and secondborn siblings of the same chronological age. *Journal of Child Language 31*, 855-873.

Bradburn, N., Sudman, S. & Wansink, B. (2004): *Asking questions: the definitive guide to questionnaire design – for market research, political polls, and social and health questionnaires.* John Wiley & Sons.

Braine, M. (1963): The ontogeny of English phrase structure: the first phase. *Language 39*, 1-13.

Brown , R: (1973): *A first language: the early stages.* Harvard University Press.

Caselli, M., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L. & Weir, J. (1995): A cross-linguistic study of early lexical development. *Cognitive Development 10*, 159-199.

Chomsky, N. (1957): *Syntactic Structures.* Mouton.

Chomsky, N. (1965): *Aspects of the theory of syntax.* MIT Press.

Dale, P. & Goodman, J. (2005): Commonality and individual differences in vocabulary growth. In M. Tomasello & D. Slobin (eds.): *Beyond Nature – Nurture: Essays in Honor of Elizabeth Bates.* Lawrence Erlbaum Associates. Chapter 2, 41-80.

Dore, J. (1983): Feeling, form, and intention in the baby's transition to language. In R. Golinkoff (ed.): *The Transition from Prelinguistic to Linguistic Communication.* Lawrence Erlbaum Associates. Chapter 10, 167-190.

Dromi, E. (1999): Early lexical development. In M. Barrett (ed.): *The Development of Language: Studies in Developmental Psychology.* Psychology Press Ltd. Chapter 5, 99-131.

Eriksson, M. & Berglund, E. (1999): Swedish early communicative development inventories: words and gestures. *First Language 19*, 55-90.

Eriksson, M. & Berglund, E. (2005): Language acquisition in Swedish children reflected by the MacArthur-Bates Communicative Development Inventories. Poster presented at the Xth International Conference for the Study of Child Language, Berlin.

Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethik, S. & Reilly, J. (1993): *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Singular Publishing Group, San Diego.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D. & Pethik, S. (1994): Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development 59(5)*, Serial no. 242.

Fenson, L., Pethick, S., Renda, C., Cox, J., Dale, P. & Reznick, J.S. (2000): Short form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics 21*, 95-115.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J.S. & Bates, E. (2007): *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* Second edition. Paul H. Brookes Publishing Co, Baltimore.

Ingram D. (1989): *First Language Acquisition: Method, Description, and Explanation.* Cambridge University Press.

Ingram, D. & Le Normand, M-T. (1996): A diary study on the acquisition of Middle French: A preliminary report of the early language acquisition of Louis XIII. *Proceedings of the 20th Annual Boston University Conference on Language Development*, 352-363.

Jakobson, R. (1962): Why "mama" and "papa"? In R. Jakobson: *Selected writings, Vol.1: Phonological studies.* Mouton, The Hague. 538-545.

Jespersen O. (1922/1964): *Language: its Nature, Development and Origin.* George Allen & Unwin Ltd, London.

Kalbfleisch, J. & Prentice, R. (2002): *The statistical analysis of failure time data.* Wiley Series in Probability and Statistics. John Wiley & Sons.

Lauritzen, L., Jørgensen, M., Olsen, S., Straarup, E. & Michaelsen, K. (2005): Maternal fish oil supplementation in lactation: effect on developmental outcome in breast-fed infants. *Nutrition Development 45*, 535-547.

Levin, J. (2005): Early lexical acquisition in 8-16-month-old Danish, American, and Italian children. Master of Arts Thesis. Department of English at the University of Southern Denmark, Odense.

Lieven, E., Pine, J. & Dresner Barnes, H. (1992): Individual differences in early vocabulary development: redefining the referential-expressive distinction. *Journal of Child Language 19*, 287-310.

MacWhinney, B. (2000): *The CHILDES project: Tools for Analyzing Talk.* Volumes 1+2. Lawrence Erlbaum Associates.

Maital, S., Dromi, E., Sagi, A. & Bornstein, M. (2000): The Hebrew Communicative Development Inventory: language specific properties and cross-linguistic generalizations. *Journal of Child Language 27*, 43-67.

Nelson, K. (1973): Structure and Strategy in Learning to Talk. *Monographs of the Society for Research in Child Development 38*, Serial no. 149.

Pavuluri, M. (2007): Parental report version of the Mood Disorder Questionnaire for adolescents has good sensitivity and specificity in diagnosing bipolar disorder in psychiatric outpatient clinics. *Evid. Based Ment. Health 10*, 9.

Piaget, J. (1952): *The child's conception of number.* Norton, New York.

Pine, J. (1992): How referential are "referential" children? Relationships between maternal-report and observational measures of vocabulary composition and usage. *Journal of Child Language 19*, 75-86.

Preyer, W. (1889): *The mind of the child.* Appleton. Translation of original German edition of 1882.

Rehfeldt, K. & Riegels, N. (2003): Sprogligt input og tidlig leksikalsk tilegnelse [Language input and early lexical acquisition]. Master of Arts thesis. Institut for Almen og Anvendt Sprogvidenskab, University of Copenhagen.

Reynell, J. & Gruber, C. (1990): *Reynell Developmental Language Scales.* U.S. edition. Western Psychological Services, Los Angeles.

Schaner-Wolles, C. (2001): Vom Umgang mit Erstspracherwerbsdaten aus Langzeit-Fallstudien [About the handling of first language acquisition data from longitudinal case studies]. In H. Gruber & F. Menz (eds): *Interdisziplinarität in der Angewandten Sprachwissenschaft: Methodenmenü oder Methodensalat?.* Peter Lang GmbH, Europäischer Verlag der Wissenschaften. 223-249.

Shore, C. (1995): *Individual Differences in Language Development.* Sage Series on Individual Differences and Development. Sage Publications.

Skinner, B. F. (1957): *Verbal Behavior.* Appleton-Century-Crofts.

Stern, C. & Stern, W. (1907): Die Kindersprache. Eine psychologische und sprachtheoretische Untersuchung. Barth (Re-print of the 4. edition in the *Wissenschaftliche Buchgesellschaft* Darmstadt, 1987).

Szagun, G., Steinbrink, C., Franik, M. & Stumper, B. (2006): Development of vocabulary and grammar in young German-speaking children assessed with a German language development inventory. *First Language 26*, 259-280.

Taine, H. (1877): On the acquisition of language by children. *Mind 2*, 252-259.

Tardif, T., Gelman, S. & Xu, F. (1999): Putting the ”noun bias” in context: a comparison of English and Mandarin. *Child Development 70*, 620-635.

Tomasello M. & Stahl, D. (2004): Sampling children’s spontaneous speech: how much is enough? *Journal of Child Language 31*, 101-121.

Vach, W. & Bleses, D. (in preparation): Association among verbs: a study of Danish children’s past tense acquisition.

Vach, W., Wehberg, S. & Bleses, D. (in preparation): The growth of the vocabulary in Danish children: Results from a longitudinal CDI-based study.

van Geert, P. & van Dijk, M. (2003): Ambiguity in child language: the problem of interobserver reliability in ambiguous observation data. *First Language 23*, 259-284.

Veneziano, E. (1981): Early language and nonverbal communication: A reassessment. *Journal of Child Language 8*, 541-563.

Vihman, M. & McCune, L. (1994): When is a word a word? *Journal of Child Language 21*, 517-542.

Vihman, M. (1996): *Phonological Development: The Origins of Language in the Child.* Blackwell Publishers.

Vygotsky, L. (1962): *Thought and language.* MIT Press. Translation of original Russian version of 1934.

Wynn, K. (1992a): Evidence against empiricist accounts of the origings of numerical knowledge. *Mind & Language 7*, 315-332.

Wynn, K. (1992b): Addition and subtraction by human infants. Letters to Nature. *Nature 358*, 749-750.

Xu, F., Spelke, E. & Goddard, S. (2005): Number sense in human infants. *Developmental Science 8*, 88-101.

Chapter 2

The Danish Longitudinal CDI study

2.1 Data description

Based on the American English MacArthur-Bates Communicative Development Inventories (CDI) (Fenson et al. 1993; Fenson et al. 1994; Fenson et al. 2007), a Danish adaptation was completed for the following two questionnaires: the Infant part "Words and gestures" was adapted to "Ord og gestikulation" in Danish, and the Toddler part "Words and sentences" to "Ord og sætninger" (cf. Andersen et al. 2006). So far, these Danish questionnaires have been applied in a number of studies (cf. Lauritzen et al. 2005; Andersen et al. 2006; Bleses et al., *in press*); in particular, large cross-sectional studies have been conducted enabling the establishment of lexical norms for Danish children (cf. Bleses et al., *submitted 1*; Bleses et al., *submitted 2*; CLEX webproject[1]).

Here we will focus on the Danish Longitudinal CDI study, which was conducted at the University of Southern Denmark (Odense) between the years 2000 and 2002 and was, in fact, one of the first data collections based on the Danish CDI. Details on recruitment, the use of additional (general) questionnaires, inclusion criteria of participants etc. can be found in Andersen et al. (2006: chapter 3). Basically, data on the 183 participating children was collected when the children were between 8 and 30 months of age – based on "Ord og gestiku-

[1]cf. chapter 1; a preliminary version can be found at *http://www.cdi-clex.org*.

Table 1. Technical description of the final Danish Longitudinal CDI dataset, "Ord og gestikulation" (Infant part), including general variables. Danish names are used for the CDI parts and categories to emphasise the Danish character of the adaptation; a translation can be found in Andersen et al. (2006). Note that variables within a block are numbered from left to right (row-wise), not column-wise.

Part	Content	Variable name	& number	Labels and data types
General variables				
	Child identification	*barn_id*		(string)
	Gender	*sex*		1 male, 2 female
	Number of siblings	*sibling*		(numbers: 0-4)
	Number of older siblings	*siborder*		(numbers: 1-5)
	Education level (parents)	*educlass_* *	*m/f*	1 Grundskole/Gymnasial U,
			(mother);	2 Kortere VU,
			(father)	3 Mellemlang VU,
				4 Lang VU
				998 Unknown,
				999 Missing
	Occupation level (parents)	*occuclass_* *	*m/f*	1 Grundniveau,
			(mother);	2 Mellemste N,
			(father)	3 Højeste N,
				998 Unknown,
				999 Missing
CDI variables				
Part 1A	Første tegn på forståelse	*i_00_*	01-03	1 yes, 2 no
Part 1B	Vendinger	*i_01_*	1-26	1 understood
Part 1C	Den første tale	*i_02_*	1,2	1 never, 2 sometimes, 3 often
Part 1 D	Ordforråd (n=410)			
	Lydeffekter og dyrelyde	*i_03_*	1-11	1 understood, 2 said
	Dyrenavne	*i_04_*	1-36	"
	Transportmidler	*i_05_*	1-10	"
	Legetøj	*i_06_*	1-8	"
	Mad og drikke	*i_07_*	1-28	"
	Tøj	*i_08_*	1-21	"
	Legemsdele	*i_09_*	1-20	"
	Små husholdingsting	*i_10_*	1-39	"
	Møbler og rum	*i_11_*	1-24	"
	Udendørsting	*i_12_*	1-14	"
	Steder	*i_13_*	1-14	"
	Mennesker	*i_14_*	1-30	"
	Leg og rutiner	*i_15_*	1-15	"
	Ord om handlinger	*i_16_*	1-53	"
	Ord, der beskriver	*i_17_*	1-36	"
	Ord om tid	*i_18_*	1-8	"
	Ord, der henviser	*i_19_*	1-11	"
	Spørgeord	*i_20_*	1-6	"
	Forholdsord og lokaliteter	*i_21_*	1-16	"
	Kvantitetsord	*i_22_*	1-10	"
Part 2A	Første kommunikative gestikulation	*i_23_*	1-12	1 not yet, 2 sometimes, 3 often
Part 2B	Leg og rutiner	*i_24_*	1-6	1 yes, 2 no
Part 2C	Handliner med ting	*i_25_*	1-17	"
Part 2D	Lader som om.. en forælder	*i_26_*	1-13	"
Part 2E	Imiterer andre voksenhandlinger	*i_27_*	1-15	"
Part 2F	Lade som om ting: eksempler	*example*		(string)

* In coding *education* and *occupation* levels of the parents, ambiguous cases were coded as 998 (Unknown); cf. Table 4.

lation" between 8-15 months and on "Ord og sætninger" from 16 months on until study end. A technical description of the final raw version of the Danish Longitudinal CDI dataset[2] is presented in Tables 1 and 2; variables from the general questionnaires ("general variables") are only included in the tables insofar as they were used in any analyses presented in this thesis.

With a few exceptions, the parts of the CDI questionnaires are ordered by increasing competence level of the children: questions about first signs of understanding, for instance, come before questions about the actual lexicon. For the final (raw) version of the dataset (Tables 1+2), corresponding checks for plausibility of the data *within* a CDI form have not been performed: it has not been checked, for example, if children with a complexity score of more than 1 also have a production score of more than zero (which is necessary). Checks for plausibility *across* CDI forms of one child, that is, over time, have not been done for the raw dataset, either.

However, for our analyses contained in the papers (chapters 3–5), the first type of plausibility checks is not relevant because we focus solely on the vocabulary checklist parts of the CDI. The second plausibility check is indirectly incorporated into the respective (single) analyses since, for example, the cumulative lexicon over time instead of the monthly data is considered in chapter 3 and 4, thus ensuring continuity and plausibility. Note that we define the "cumulative (productive) lexicon" of a child at the beginning of at month 20, say, as covering all words or items of the checklist which have been reported as "understood and said" at least once before month 20; this coincides with counting all items which are reported as said in the first month 8, and then successively adding all new items, which occur for the first time in the subsequent months 9 up to 19.

In the following we will first have a closer look at validity and reliability of the Danish CDI in a longitudinal design, before we explore the longitudinal data in terms of test-retest stability or predictive values to gain additional information on the Danish CDI in general.

[2]This version is available in Stata^{TM}, copyright by StataCorp LP.

Table 2. Technical description of the final Danish Longitudinal CDI dataset, "Ord og sætninger" (Toddler part). Danish names are used for the CDI parts and categories to emphasise the Danish character of the adaptation; a translation can be found in Andersen et al. (2006). Note that variables within a block are numbered from left to right (row-wise), not column-wise.

Part	Content	Variable name	& number	Labels and data types
CDI variables				
Part 1A	Ordforråd (n=725)			
	Lydeffekter og dyrelyde	*i_01_*	1-12	1 said
	Dyrenavne	*i_02_*	1-43	"
	Transportmidler	*i_03_*	1-14	"
	Legetøj	*i_04_*	1-18	"
	Mad og drikke	*i_05_*	1-68	"
	Tøj	*i_06_*	1-30	"
	Legemsdele	*i_07_*	1-28	"
	Små husholdingsting	*i_08_*	1-50	"
	Møbler og rum	*i_09_*	1-33	"
	Udendørsting	*i_10_*	1-31	"
	Steder	*i_11_*	1-22	"
	Mennesker	*i_12_*	1-40	"
	Leg og rutiner	*i_13_*	1-27	"
	Ord om handlinger	*i_14_*	1-103	"
	Ord, der beskriver	*i_15_*	1-63	"
	Ord om tid	*i_16_*	1-15	"
	Ord, der henviser	*i_17_*	1-31	"
	Spørgeord	*i_18_*	1-7	"
	Forholdsord og lokaliteter	*i_19_*	1-41	"
	Kvantitetsord	*i_20_*	1-21	"
	Hjælpeudsagnsord	*i_21_*	1-21	"
	Forbinderord	*i_22_*	1-7	"
Part 1B	Hvordan børn bruger ord	*i_23_*	1-5	1 yes, 2 no
Part 2A	Ords endelser (del 1)	*i_24_*	1-3	1 not yet, 2 sometimes, 3 often
Part 2B	Ordformer	*i_25_*	1-29	1 said
Part 2C	Ords endelser (del 2)	*i_26_*	1-62	1 said
Part 2D	Spørgsmål vedr. kombination af ord	*i_27_1*		1 not yet, 2 sometimes, 3 often
	Eksempler	*example_*	1-3	(strings)
Part 2E	Kompleksitet	*i_28_*	1-33	(1,2: 2=complex version)

2.2 Validity and reliability of the Danish CDI in a longitudinal design

2.2.1 Introduction

Validity and reliability of the original CDI have been investigated extensively. For example, in Fenson et al. (1993), reliability is evaluated investigating internal consistency, test-retest correlation, and standard errors of the mean. Further, validity is discussed with regard to various aspects – face, content, and convergent validity in terms of concurrent and predictive validity – where the latter part is mostly supported by correlational arguments. Overall, the authors conclude that "There is now a large body of evidence supporting the reliability, validity, clinical utility, and research potential of the MacArthur Communicative Development Inventories" (Fenson et al. 1993: 77; cf. also Fenson et al. 2007: 114).

Results of some investigations into validity and reliability of the Danish adaptation of the CDI have been reported in Bleses et al. (*submitted 1*). The authors conclude that "a) most common words and a substantial part of the less common words used spontaneously by Danish children are included in the vocabulary lists, b) the vocabulary development of Danish children as measured by the CDI seems to correlate adequately with the growth of word types in spontaneous speech as measured in four Danish children's spontaneous speech productions, and c) values of internal consistency of four different scales of the Danish CDI were found satisfactory" (page 12, submitted version of 22/04/2007).

In the following we will investigate and discuss the possible implications of a longitudinal design on interpretability and comparability of study results by looking at the following topics:

- **Internal consistency.** Are the results on internal consistency based on the longitudinal study similar to the cross-sectional results?
- **Sampling.** Are parents who decided to participate in the longitudinal study essentially different to other parents in Denmark? Or more specifically, how comparable is the demographic composition of the longitudinal sample to that of the general population?

- **Compliance.** Who drops out of the study and how are missing values distributed (where a missing value is a whole CDI form which has not been filled in or sent at a specific month)?

- **Parental form-completing behaviour.** Do longitudinal parents[3], for example, employ strategies such as "copy and add" in filling in the questionnaires from month to month? If they did, the comparability of longitudinal results, evaluated at each month separately, to cross-sectional results could be seriously affected (see "pointwise results" below).

- **Pointwise results.** Are the overall longitudinal results – as measured by some selected sum scores (pointwise) for each month – comparable to the cross-sectional results? Here, we also look at the placement of the longitudinal children with respect to the lexical norms in Danish.

Table 3. Cronbach's standardised alpha values

	LS	CSS 8–30	CSS 8–36	ACSS*
N of children:	183	4400	6112	1789
(Infant + Toddler):		(1537+2863)	(2398+3714)	(659+1130)
Infant Part, Word Perception	.9785	.9702	.9819	.95
Infant Part, Word Production	.9339	.9170	.9733	.96
Infant Part, Gestures	.9061	.8909	.9112	.39 (.88*)
Toddler Part, Word Production	.9910	.9893	.9912	.96
Toddler Part, Complexity**	.9739	.9575	.9706	.95

* Available numbers taken from Fenson et al. (1993: 67-68); for Infant Part, Gestures, the value reported in Fenson et al. (2007: 100), is included in parentheses (based on an updated American sample of 2550=1089+1461 children).

** Whereas the American value is based on the three subscales "bound morphemes", "functor words", and "complex sentences", the Danish values are based on all 33 separate items without further grouping.

[3] To avoid the lengthy term "parents of children participating in the longitudinal study", the abbreviations "longitudinal parents", and correspondingly, "cross-sectional parents" will be used in the following. The expressions "longitudinal children" and "cross-sectional children" will also be used.

2.2.2 Internal consistency

To check for internal consistency we computed Cronbach's alpha (e.g., Bland & Altman 1997) for the following (standardised) scores: word perception, word production, and gestures in the Infant part, and word production and complexity in the Toddler part (Table 3). In analogy to published results (Fenson et al. 1993; Bleses et al., *submitted 1*), we used the corresponding scores in the sub-categories (subscales) as "items" for all scores, except for complexity. The resulting values for the longitudinal study (LS) are all on a very high level, spanning from .91 to .99 for gestures and Toddler word production, respectively, and are very similar to those based on the cross-sectional study (CSS), both within the matched age range 8-30 months (CSS 8–30) and within the extended age range of the norming data where the Infant data stretches from 8-20 months and the Toddler data from 16-36 months (CSS 8–36; cf. Bleses et al., *submitted 1*). The results also agree well with alpha values published for the American norming CDI data (ACSS; Fenson et al. 1993: 67-68), except for the surprisingly low value of .39 attached to the American gestures score (which was corrected to .88 in the second edition, see Fenson et al. 2007: 100). Overall, the internal consistency of the longitudinal data corresponds well to that of the cross-sectional Danish CDI findings.

2.2.3 The longitudinal sample: Are families different?

There is no doubt that people who voluntarily take part in scientific studies *are* special; even more are people who spend a considerable amount of time and energy in taking part in a 23-months lasting longitudinal CDI study. None the less, results of (voluntary) cross-sectional and longitudinal studies may be transferable to, for example, the general population if the self-election of participants does not influence the results too much – or influences them in a known way. In Table 4 we compare the longitudinal sample (LS) to the samples of the cross-sectional Danish norming studies (CSS) and to two Danish populations, the overall general population (GP) as well as the general child-family population (CFP): we look at the distributions of *gender* and *sibling status* of the child as well as *education* and *occupation* level of the parents to see in what ways the populations differ, if at all.

Not surprisingly, there is not much difference between all four populations with respect to *gender* and not much difference between the three child-family populations with respect to

Table 4. Composition (in %) of the longitudinal sample (LS) compared to an age-matched sample of the cross-sectional norming studies (CSS; Infant part 8-15; Toddler part 16-30), the general child-family population in Denmark (CFP) and the general population in Denmark (GP). Numbers for CFP and GP are taken from Andersen et al. (2006: 26). To make percentages comparable, missing values in *education* but not *occupation* are excluded in calculating them. Note that numbers for CSS are not directly comparable to corresponding numbers reported in Andersen et al. (2006), since there the whole norming sample from 8-36 months has been considered and slightly different coding principles for *education* and *occupation* have been employed.

	LS (N)	LS (%)	CSS 8–30 (%)	CFP (%)	GP (%)
	n=183		n=4400		
Gender					
Girls	n=95	52	50	49	49
Boys	n=88	48	50	51	51
Sibling status at study start					
Only child	n=64	35	40	40	12
At least 1 sibling	n=119	65	60	60	88
Education *					
Basic Education	n=14	8	4	27	34
Short further Education	n=85	47	46	53	46
Medium further Education	n=39	22	28	11	14
Long further Education	n=43	24	22	9	6
Missing/Unknown/Other	(n=2)		(n=197)	(n=0)	(n=0)
Occupation **					
Basic	n=24	13	8	41	46
Medium	n=70	38	39	16	19
High	n=53	29	27	19	24
Missing/Unknown/Other	n=36	20	26	23	10

* **Education**: Highest education of mother/father; categorised according to classification system established by *Statistics Denmark* (*www.dst.dk*), which follows international guidelines

** **Occupation**: Highest occupation of mother/father; categorised according to classification system established by *Statistics Denmark*, which is based upon the international standard *ISCO* *sibling status*. Regarding *education* and *occupation* level of the parents, there are distinct differences between the CDI study-based and the population-based numbers, while the composition of the two CDI study samples (longitudinal and cross-sectional) as well as the two Danish populations (overall and child-family population) agree quite well pairwise. Roughly, there are 20-30% more families with a basic education in the population-based than in the CDI studies, and in return, 10-15% more with a medium and 15-20% more with a long further education in the CDI studies than in the population. With regard to *occupation* level, mainly the percentages of basic and medium classes appear switched: in the CDI studies, there are only 13 and 8% of basic versus 38 and 39% of medium level occupation, whereas in the Danish population(s), there are 41 and 46% of basic versus only 16 and 19% of medium level.

To summarise, the impact of a longitudinal design on self-election of families and therefore, on the (initial) composition of the study sample, seems rather small: the composition of the Danish longitudinal CDI study sample is quite similar to that of the Danish CDI norming studies in the corresponding age range, at least if judged by the demographic criteria *gender*, *sibling status*, and *education* and *occupation* level of the parents. However, note that both CDI studies included more families where the parents have a longer education and/or a higher-ranked occupation than a random sample of the (child-family) population on average would have. This has to be taken into account whenever transferring study results to the general (child-family) population.

2.2.4 A description of compliance: missing values and drop-outs

In this section, we study sample selection over time by looking at missing values (whole CDI forms not sent in, at a given month) and drop-outs (families who discontinue study participation before their child is 30 months old). Here, for example, we address the question whether all poor-performing children drop out at about 26-27 months, leaving only the top-level children in the longitudinal sample.

As presented in Table 5, almost three quarters (73.2%) of all 183 participating children had at most 3 missing monthly observations out of 23 possible observations (March 2000 – January 2002), and almost half of the children (47%) contributed at all possible time points. A little

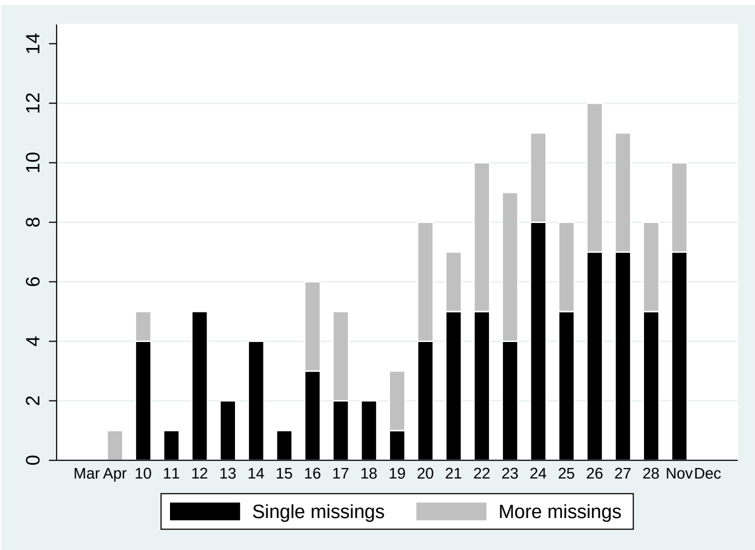
Table 5. Overview over general missing pattern

Age at last observation			Number of monthly observations per child		
Age	N	%	Counts	N	%
13	1*	0.5	6	1*	0.5
20 – 24	19	10.4	12-13	4	2.2
25 – 27	22	12.0	14-19	44	24.0
28 – 29	23	12.6	20-22	48	26.2
30	118	64.5	23	86	47.0

* This one child has been excluded for the analyses in chapters 3 and 4.

more than three quarters (77.1%) of the children continued at least until month 27, and 64.5% completed the study at 30 months (though possibly with missing values during study course). In the following we first investigate the number of missing CDI forms per month, that is, the number of children who miss one or more monthly observations in a row (not sending in CDI forms) before continuing with the study. We define a missing CDI form as a "single missing value" if a family sent in a questionnaire the month before and the month after. If

Figure 1. The distribution of "single" and "other" missing values over time



a family missed more than one monthly observation in a row, but continued afterwards with sending in questionnaires, this is considered an "other missing value". Figure 1 shows the distribution of missing values over time: overall, there are relatively few missings, although the total number – not surprisingly – increases with study duration; the largest numbers of missings are in July and September months, indicating some seasonal influence.

We proceed by taking a closer look at the drop-out children, that is, at the 65 children who discontinued the study participation before 30 months of age (Table 6). Not surprisingly, the number of drop-outs increases with time (months), with the largest drop-out rate – almost 10% of all children still participating in the study – just one month before study end. Among all drop-out children, there are slightly more girls than boys (57% versus 43%). The hypothesis that drop-outs are mostly poor-performing children is not supported at all – on the contrary, children with a larger than average lexicon seem more prone to terminate study participation early.

2.2.5 Parental form-completing behaviour

Here, we describe the parents' behaviour over time in filling in the CDI forms; in particular, we consider the possibility that parents develop form-completing strategies over time to help them cope with their task, for example, copying CDI forms and just adding new words at the next month. Employing this strategy would imply, for example, production scores reflecting cumulative lexica instead of momentary glimpses at a lexicon present at a certain month. As a consequence, the resulting longitudinal CDI data might not be directly comparable to cross-sectional CDI data – and that would be an undesirable effect of the longitudinal design. Specifically, we look into the parents behaviour at the following three levels: 1) lexicon (production) level: how large is the difference between a cumulative lexicon and a monthly lexicon, that is, are there actually parents using a "copy and add" strategy?, 2) category level: is there a general difference between cross-sectional and longitudinal parents in reporting items from specific CDI categories?, and 3) single item level: how are inversions – a word (or item) goes from "understood and said" in one month back to not or only "understood" in the next – distributed among children and words?

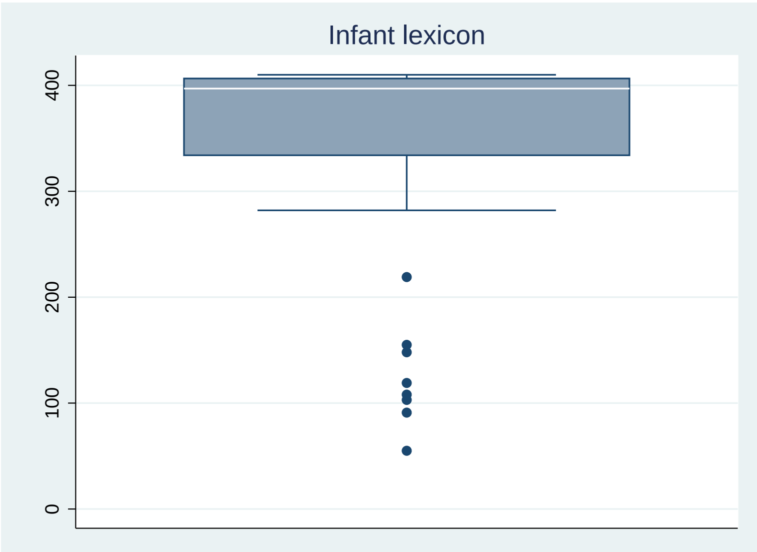
Table 6. Drop-outs: information about the children dropping out of the study is presented at their last observable month. Percentages are based either on all 183 children (%1) or on all children still in the study (%2). Drop-out children are categorised based on production score quartiles of the cross-sectional study and of the longitudinal study (in parentheses) at each month (cf. section 2.2.6 below; Figure 7).

Months	N	%1	%2	N of girls	Number of drop-out children in performance categories*			
					low	middle-low	middle-high	high
13	1	0.5	0.5	1				1
20	1	0.5	0.5	0				1
21	4	2.2	2.2	4		1 (0)	1 (2)	2
22	1	0.5	0.6	0				1
23	7	3.8	4.0	4		1	2	4
24	6	3.3	3.5	4	2	1	1	2
25	9	4.9	5.5	5		4	3 (2)	2 (3)
26	6	3.3	3.9	5	3 (2)	0 (1)	1	2
27	7	3.8	4.7	4		3 (4)	2	2 (1)
28	10	5.5	7.1	5	2	1 (2)	1 (3)	6 (3)
29	13	7.1	9.9	5	4	4	0 (3)	5 (2)
	65	35.4 (%)		37	11 (10)	15 (17)	11 (16)	28 (22)

* Performance categories are: low = smaller than 25% quantile, middle-low = smaller than median and larger than/equal to 25% quantile, middle-high = larger than median and smaller than/equal to 75% quantile, high = larger than 75% quantile.

Ad 1) Lexicon level: are there parents who actually use a "copy and add" strategy in filling in the CDI forms?

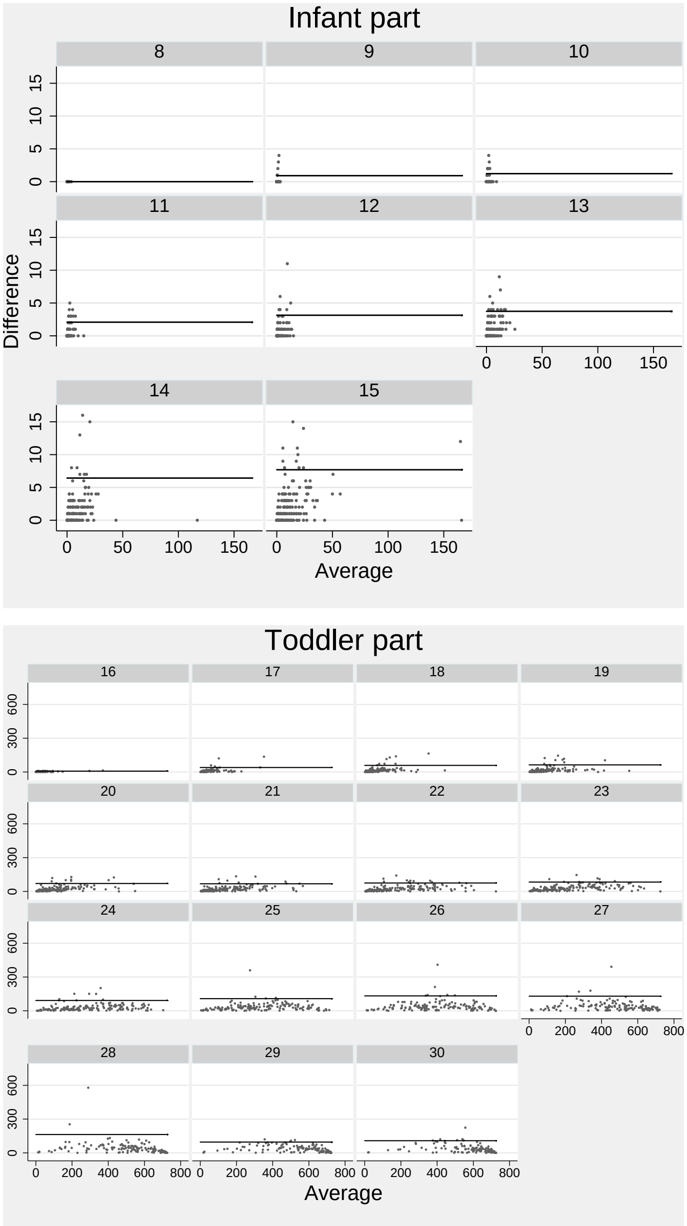
Figure 2. The distribution of the cumulative (production) Infant lexicon (max. 410) in 56 child-months pairs where there are zero inversions in the next month, which is observed



First, we investigate whether there are families who always "copy and add". In the Infant part, there are 48 children (out of 183) whose cumulative lexicon[4] coincides with the noted single-month lexica *at each observed month*. But for 284 out of these 376 (= 48 children x 8 months - missing months) child-month pairs, the lexicon actually equals zero. In the Toddler part, there are at most six (out of 15 possible) months per child where there is total agreement between the cumulative lexicon and the single-month lexicon; and there are only three children with more than four observations with such an agreement (5 agreements: 1 child; 6 agreements: 2 children). Therefore, there is no indication that some families strictly use "copy and add".

[4] As stated above, the "cumulative (productive) lexicon" of a child at the beginning of at month 20, say, is defined as follows: the sum of words or items of the vocabulary checklist which have been reported as "understood and said" at least once before month 20.

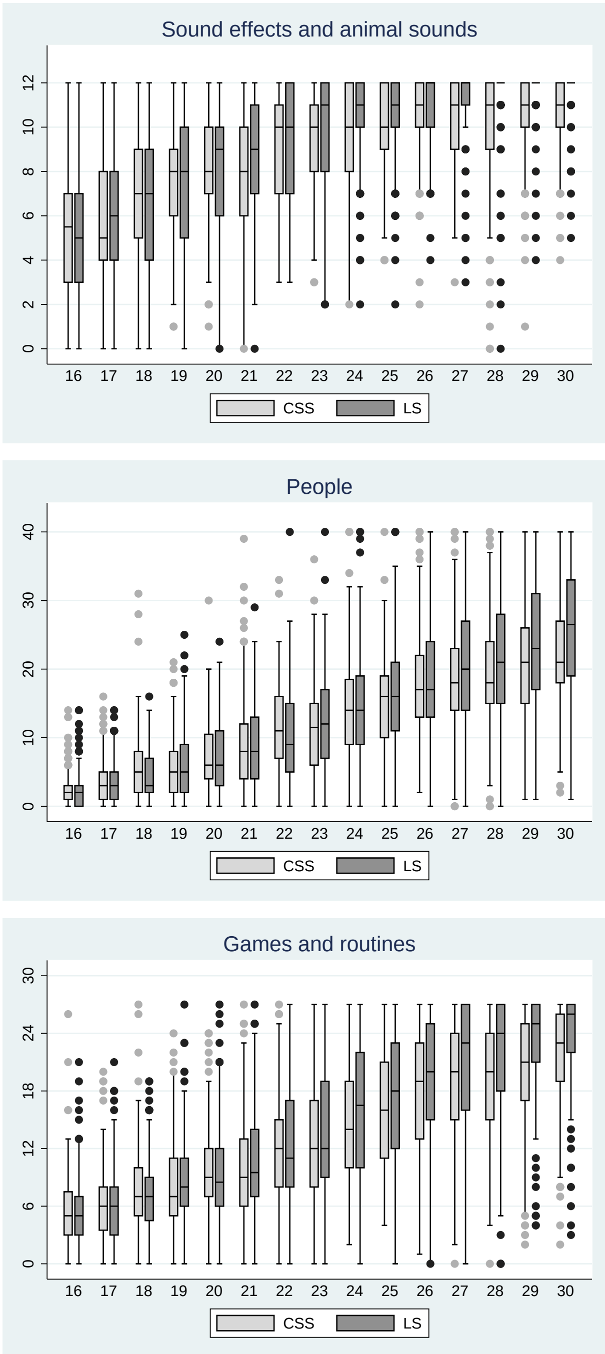
Figure 3. Bland and Altman plots (Altman & Bland 1983; Bland & Altman 1986): Average of cumulative lexicon and single-month score (x-axis) versus difference between both (y-axis). The line represents the average plus (estimated) 2 standard deviations (under normality, approximately 95% fall below this line).



But are there families who employ this strategy from a certain level of their child on, since "copy and add" makes most sense when the children are actually adding many new words to their lexicon each month – which is, for most of the children, first happening in the Toddler part phase? To investigate this, we consider only months where the children have a cumulative lexicon of at least 50 words (out of 410 potential Infant words) and calculate for each child-month pair the number of inversions: how many items go from "said" the month before back to "not understood/only understood" in the present month. Then, zero inversions are equivalent to "just adding new words", implying the possibility that parents actually make use of "copy and add". There are 178 children who reach a cumulative lexicon of 50 Infant items at one point during the study, covering 1652 child-month pairs. The number of inversions (in these pairs) ranges from zero to 321, where there are 56 child-month pairs (out of 1652) with no inversions, distributed among 43 different children. 32 children have zero inversions in just one single month, 9 children in two months, 2 children in three months, and no children have zero inversions in more than three months. Looking at the distribution of the cumulative lexicon at the months with zero inversions (Figure 2), we see that it is mostly children with a large cumulative lexicon, near the ceiling of 410 Infant words, where parents tend to just add new words. This indicates that, overall, the longitudinal parents apply "copy and add" only seldomly, and if they do, their child has reached a close-to-the-ceiling cumulative lexicon.

If analyses on the single-month lexica are to be performed on the longitudinal data (which we generally do not attempt, cf. chapters 3 - 5), the information contained in Figure 3, showing the differences between the cumulative lexicon and the single-month lexica versus their average at each month, could in principle be used to, for example, identify possible outliers. Some of the outlier values visible in Figure 3, especially in the Toddler part, might be explained by a change in person who completed the form (cf. above, where the maximum number of inversions is reported as 321). Unfortunately, information on who of the parents or caregivers actually completed the CDI form at a specific month is missing in the Danish Longitudinal CDI study.

Figure 4. Score distributions within single lexicon categories in cross-sectional (CSS) and longitudinal (LS) Danish CDI studies over time



Ad 2) Category level: is there a general difference between cross-sectional and longitudinal parents in reporting items from specific CDI categories?

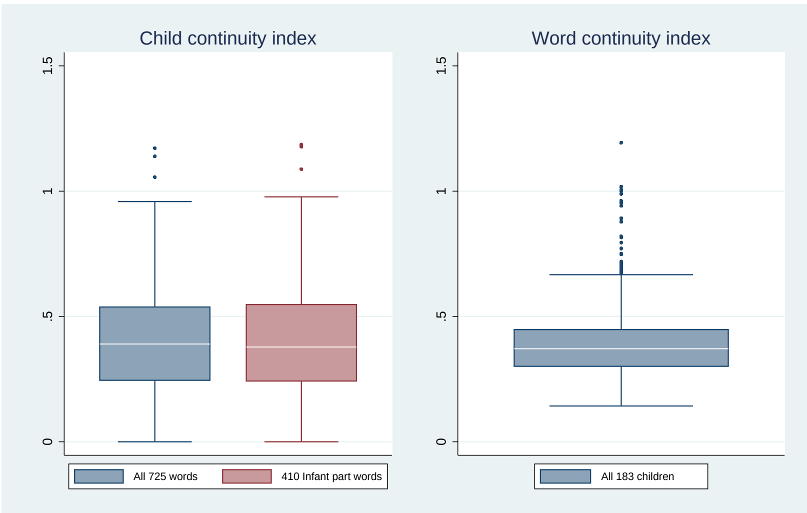
Figure 4 shows the differences between longitudinal and cross-sectional scores for three exemplary categories: *Sound effects and animal sounds*, *People*, and *Games and routines*. We see that, towards the end of the study period, the longitudinal scores (quartiles) lie slightly above the cross-sectional scores for all three categories. This finding could be explained by the following (extreme) hypothesis: since in all three categories, (very) early emerging items are predominantly placed, the differences might be due to the fact that longitudinal parents at the end of the study (still) mark all of the items which their child used earlier on, whereas cross-sectional parents only evaluate their child's *current* productive repertoire and never include these earlier used items in their reporting. However, even though there might be general tendency for this behaviour, the hypothesis in its extreme form is not supported by Figure 4, since the longitudinal scores are only slightly above the cross-sectional results. At any rate, cross-sectional parents *do* mark items as, for example, *vov (dog sound)*, which 98% of the 190 28-month old children in the Danish CDI norming study are reported to use.

Ad 3) Single item level: how are inversions distributed among children and words?

The analysis of overall production scores as presented above under 1) (lexicon level) is a rough measure of continuity and stability of (production) information over time. In the following we look at the stability of single items and base a more detailed version of continuity measure on the number of inversions of single items (words): for each child and each item, we count the number of times, the item has been marked as "said" in one month, but as "not said" or only "understood" in the following month. A small number of inversions would indicate "stable" words (or children whose parents' reporting is stable), and a larger number of inversions would mark a more "unstable" word (or child). In principle, there could be at most 11 inversions (for an Infant item), if a child produced a specific item in month 8, then went back to "understood" in month 9, alternating this until month 29: this results in 11 times of "switching back". In our data, we find that the total number of inversions per child and item ranges between zero and six inversions.

To get a better overview, we begin by looking at words which have at least one inversion for each child (that is, for each child who produced a word at least once during study course), to point out relatively unstable words. Here, we find that 149 of all 725 words (21%) have at least one inversion for each of the 183 children. Lowering the demand to "at least one inversion for *almost any* child" (instead of "at least one inversion for *each* child"), we identify almost all words as unstable (93% of all words have at least one inversion for more than 177 children, for example). There are only four words where less than 171 children have at least one inversion: *aah aahh (ups/being-surprised sound)* as the most stable word where for (only) 142 children at least one inversion is reported; further, there are *traktor (tractor)* for 162 children, *hjemmesko/sutsko (slippers)* for 168 children, and *farvel (bye-bye)* for 170 children. Since this criterium "words which have at least one inversion for each child" is rather crude, identifying almost any item as unstable if we modify "for each child" slightly, we will in the following consider a different, more sensitive criterium, the *word continuity index*.

Figure 5. Continuity indices: relative number of inversions per child across words (left) and per word/item across children (right)



The *word continuity index* is defined as the total number of inversions divided by the total number of children who said this word at least once during study course: a large index[5] points to unstable items, whereas stable items tend to a smaller index. The right-hand side of Figure 5 shows the distribution of the *word continuity index*. It is noticeable that there are no words with no inversions at all, and that there are many outliers at the upper end around 1, for example, *borte tit/titte bøh (peekaboo/hiding game)*, *hvad (what?)*, *her (here)*, *grrrr (lion sound)*, *det (this/that; the)* (both pronoun and article), *mere (more)*, and *der (there)* which have continuity indices larger than 0.9; their inversion maxima (number of inversions per child) range from 4–6. These words are weak, unstable items in the sense that they are inverted often as opposed to, for example, *arm* and *cola (coca cola)* which are the only two words with indices below 0.15.

To obtain a corresponding index for each child (cf. lexicon level, above), we define the *child continuity index* as the total number of inversions divided by the total number of words the child said at least once during study course. Note that the denominator simply counts the number of words learnt and does not further differentiate between, say, words learnt in month 25 or words already learnt in month 8, although the latter are much longer at risk for an inversion. The left-hand side of Figure 5 shows the distribution of this index considering all words (left) as well as restricted to only Infant part words (right). The distribution looks fairly symmetric (given that the lower bound is zero). Only three children have an index slightly larger than 1; their inversion maxima range from 3–4. Nine children have very small indices below 0.05, whereof only one child – with a cumulative productive lexicon of 10 words at end of study – has no inversions at all.

Summary

On the lexicon level, there is only one family who employs a strict "just adding new words" strategy resulting in equality of the cumulative and the single-months lexica at all observed months, but this is the one family dropping out at month 14 when their child has a lexicon of 10 words (cf. Table 5. Not surprisingly, months with zero inversions are frequent when the child has no or only few words in her vocabulary. Also at the upper level, that is, when

[5] Note that the index can become larger than one.

children have a large cumulative lexicon close to the ceiling of 410 (Infant) words, we find some incidents of zero inversions, meaning that parents tend to add new words or, possibly, apply "copy and add" when their child knows many words. However, overall, there are only few months with zero inversions, if the child's cumulative lexicon exceeds 50 words. As a consequence, the notion that parents note the cumulative lexicon rather than remember and evaluate the child's lexicon at each month cannot be supported. This indicates that reporting behaviour on the lexicon level agrees well for longitudinal and cross-sectional parents.

With respect to the category level, there is a slight longitudinal advantage for some exemplary early categories, at least towards the end of the study period. An explanation might be, that longitudinal parents tend to mark items which their child used to say, whereas cross-sectional parents evaluate their child's current lexicon, not covering former uses. However, this tendency is not very distinct: it is not the case that cross-sectional parents never mark these early items.

Not surprisingly, results on the single item level agree well with results on the lexicon level. However, here, the positive finding of "no general strategic difference between longitudinal and cross-sectional behaviour" gets the more negative interpretation of "longitudinal parents do not report in a stable way": there are many inversions between months. Regarding the child index, the results here confirm the results from the lexicon level. With respect to single words, the word continuity index indicates differences between words with respect to parents' reporting behaviour. Since this has no direct impact on the comparability of longitudinal to cross-sectional results, it will be discussed elsewhere (cf. section 2.4.1).

2.2.6 Comparability of pointwise overall results

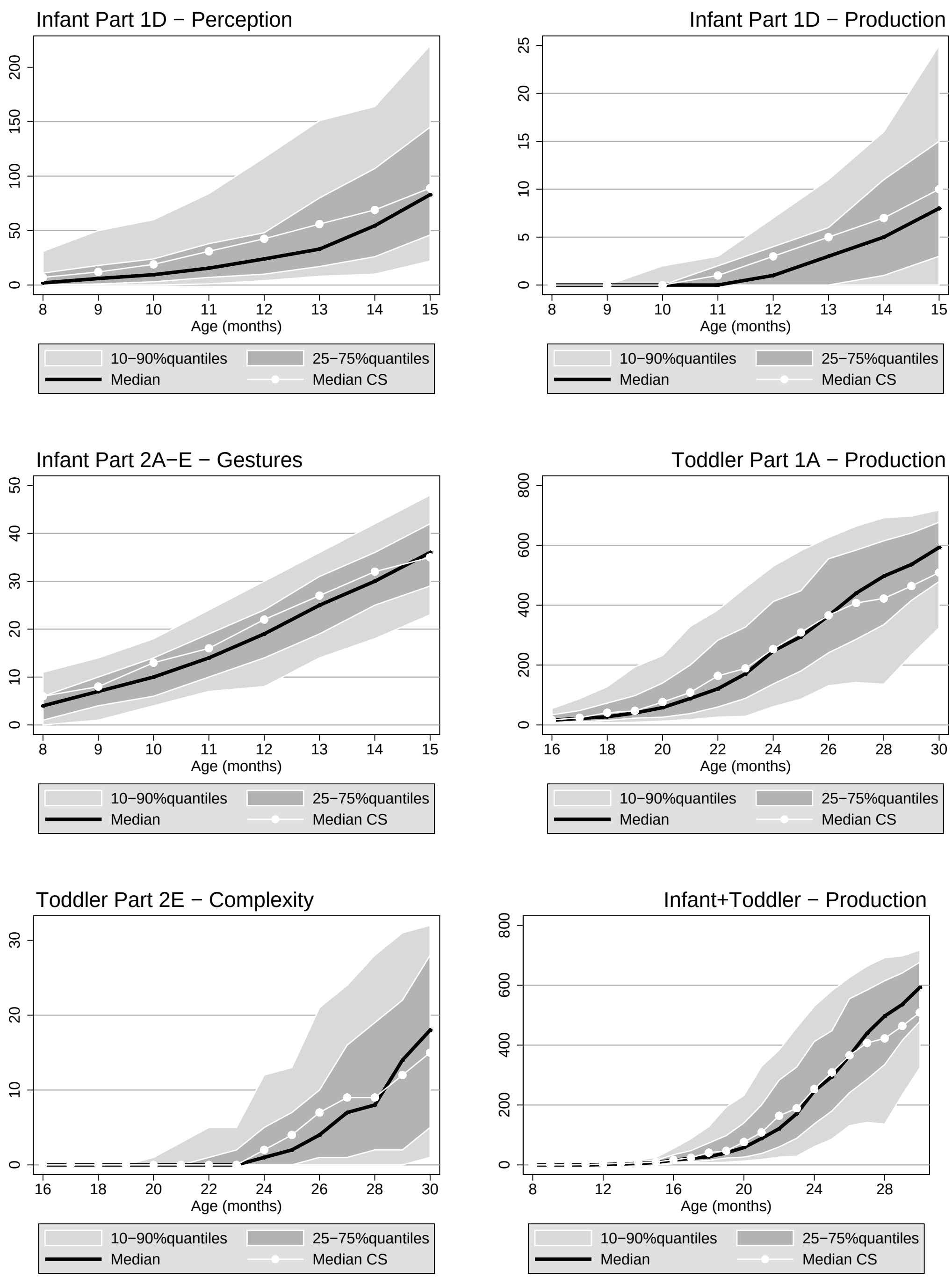
In Figure 6, the distribution of several selected sum scores for both Infant and Toddler part of the longitudinal study data is shown (shaded areas). The calculations were done pointwise – in a cross-sectional-like fashion – for each month separately; each child contributed only with her observed data for each given month, that is, the longitudinal character of the data was not taken into account and, consequently, possible imputations of missing monthly data were not considered. To facilitate comparison, the median scores based on the CDI norming studies (CS) are added as white connected dots.

For all scores, the middle area (25-75% percentiles) of the longitudinal scores contains the "normed" medians (median scores of the CDI norming studies), indicating a reasonably comparable average development over time – at least, the longitudinal sample is not extremely (high or low) positioned with respect to the score distributions of the CDI norming study. From the beginning until the middle of the study period, the median of the longitudinal sample tends to lie slightly below the normed median for all scores; for the perception and production scores, for example, the difference between the medians is 23, 14.5, and 6 words (perception) or a constant 2 words (production) for the last 3 months of the Infant part (13-15 months). Despite a slower start, the longitudinal children catch up on their cross-sectional peers, and even outperform them in the last months of the study period, this being visible for scores covering the Toddler part, that is, production and complexity, where the medians are crossing at around 26-27 and 28-29 months of age, respectively.

For the production score, this is further illustrated in Figure 7 where children participating in the longitudinal study are classified with respect to the overall Danish lexical norms each month into the following categories: comparatively low score (smaller than normed 25% quantile), lower middle score (smaller than normed median and larger than or equal to normed 25% quantile), equal to normed median[6], upper middle (larger than normed median and smaller than or equal to normed 75% quantile), and comparatively high score (larger than normed 75% quantile). If the quantiles of the distributions coincide, one would expect a picture similar to the bars between 14 and 26 months – before, production has rarely started (for months 8-10, the medians are zero, going up to 5 (cross-sectional) and 3 (longitudinal) words in month 13, respectively), and afterwards, children from the longitudinal sample produce notably more words: the median difference lies between 72 and 84 words for the last three months of the Toddler part (28-30 months). In the last month, the median production score in the longitudinal study equals 593 words and even exceeds the normed 75% quantile of 567 words.

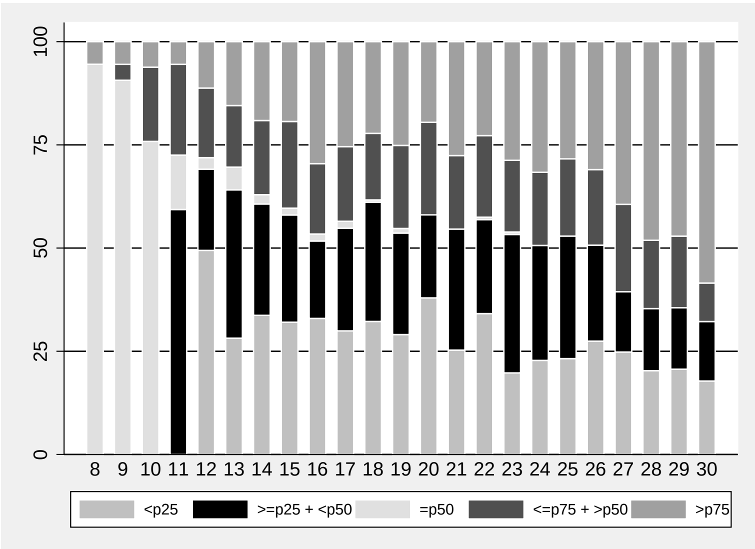
[6] This category was included to account for the relatively large number of children, especially in the beginning, who belong to this category.

Figure 6. Distribution of overall scores for the longitudinal sample (LS). Medians of the cross-sectional CDI study (CS) are included for comparison.



To summarise, the overall line of development for the presented sum scores is quite similar for the longitudinal and cross-sectional data. In the beginning, however, the medians of the longitudinal study lie slightly below the normed medians for all scores, until the medians of the Toddler scores cross towards the end of the study period. Afterwards, the medians of the longitudinal study lie considerably above the normed medians, but, at least until month 30, still below the corresponding normed 75% quantiles.

Figure 7. Children from the longitudinal study (n=183) are categorised based on their production score with respect to the Danish lexical norms established in Bleses et al., *in press*). Presented are relative frequencies based on the total number of children in each month.



Since we want to be sure that the advantage of the longitudinal children towards the end of the study period is real, we discuss in the following, whether the (long-term) time effect we see on overall Toddler scores in the longitudinal study might be artificial, that is, caused by something else than well-performing longitudinal children. A first possibility is the dropping-out of poor-performing children towards study end, biasing the longitudinal sample towards higher scores. However, making use of the results on drop-outs from section 2.2.4 above, we find no indication that the better performance of the longitudinal children (with respect to the sum scores at the end of the study period) is due to a skewed self-selection process over time: it is *not* only the top-scorers who stay in the study.

Second, the difference might be caused by the fact that the (longitudinal) parents' still report early word forms contained in the early categories as *Sound effects and animal sounds* at an age where most (cross-sectional) parents refuse to check them off, since their children are not actually using them (anymore). This implicates, that the difference we see between longitudinal and cross-sectional results is mainly based on the *early* categories. And indeed, relying on results from section 2.2.5, longitudinal scores lie (on average) above cross-sectional scores for three early categories. However, early words are also marked frequently by cross-sectional parents, and the differences visible in the early categories are too small to be solely responsible for the late advantage of longitudinal children, which, on the contrary, seems evenly distributed among all CDI categories.

Third, families do not employ a strict "copy and add" strategy, nor are there families who apply "copy and add" frequently during study course either, indicating that parents rather evaluate their child's vocabulary each month anew instead of just copying (or remembering) what they filled in the month before. This behaviour is comparable to what parents do in a cross-sectional design. Therefore, we conclude that the advantage of longitudinal children at the end of the study period is not caused by any of these potential reasons, but is a real advantage.

Speculating about the causes for this specific difference, there are the possibilities that 1) randomly, more children with a steeper than average development profile (at least with respect to production and complexity score) are included in the longitudinal sample, and 2), that participation in the longitudinal study itself acted as an intervention: the children might have got more (language-oriented) stimulation, maybe even training focused on specific CDI words; the parents themselves might have become more trained, experienced, and more careful and attentive observers (or, vice versa, more careless and sloppy?), maybe also concentrating on some specific CDI items. Additional data would be needed to differentiate between these potential causes.

2.2.7 Discussion and conclusion

Overall, we found that the internal consistency is satisfactory, the (initial) composition of the longitudinal sample is quite comparable to that of the cross-sectional sample – according to some selected criteria – and the distributions of some generally used sum scores agree (pointwise) relatively well at each month. Therefore, we conclude that there is a large initial as well as pointwise similarity between the Danish longitudinal and cross-sectional (norming) CDI studies: the Danish CDI works well within a longitudinal study design, even on a monthly basis. However, there are some time effects which should not be neglected when analysing and interpreting results from a longitudinal version of a CDI study. For example, in our longitudinal CDI study, the number of missing values per months increases over time, children stop participating in the study[7] (perhaps for a reason depending on study results), parental reporting changes more or less from month to month, and production and complexity growth curves are steeper than the corresponding (normed) average towards the end of the study period. Therefore, especially the transfer of results on predictive values (cf. section 2.3 below) or developmental profiles – which are results fully exploiting the original strength of longitudinal data – to a general Danish (child-family) population should be done very cautiously. An explanation for this higher-than-average performance might be that children and/or parents are (slowly) trained to the task – to the specific CDI task or the language acquisition task in general[8]. Perhaps, participation in a longitudinal CDI study should be mandatory since it might improve children's productive and parents' perceptive communicative competences and result in a more successful dialog between parents and children. At any rate, it would be very interesting to re-examine the longitudinal children today to see how their current language competences match a Danish norm.

[7]Although surprisingly many parents persevere until the end, given that there was no reward money or other incentives.

[8]Here, self-election of the longitudinal sample might play a role as well. With respect to results based on single-subject studies, Bennett-Kastor (1988) stated that "[these] are often based on the researcher's own offspring who, being the children of academicians, may indeed be more verbally advanced than the norm" (page 50).

2.3 The Danish CDI over time: some time dependent characteristics

In this section we present several characteristics for the Danish CDI which are only possible to determine with longitudinal data. In detail, we look first at test-retest correlation between sum scores as a usual measure of reliability and stability of the data[9] and, second, we estimate predictive values of the production sum score to get an idea about the diagnostic potential of an early CDI measurement. Note that results may be transferred to general applications of the Danish CDI only under the reservations expressed above.

Test-retest correlation

Following the reasoning in Fenson et al. (1993) that "a considerable degree of stability would certainly be expected over a period of 6 weeks or less" (page 68), we can use our monthly longitudinal data to calculate a correlation coefficient between adjacent score measurements to establish a test-retest reliability figure. The results on the longitudinal data are presented in Table 7, together with American results taken from Fenson et al. (1993: 68) as a reference. For the word perception and gestures scores of the Infant part as well as for the Toddler production score, there is a good agreement between the Danish and the American findings. In Infant word production, however, the overall Danish correlation coefficients (LS 1 and LS 2) are .77 and .72 and, consequently, a bit smaller than the reported American minimum of .80. In addition, the Danish results show an increasing strength of correlation over time, beginning with an estimated correlation of .08 between measurements of the first two months, where most children have not yet begun with word production, up to a correlation coefficient of .78 at the last two months in the Infant part (months 14 and 15), whereas the published American figures indicate no such trend[10].

[9] In principle, the method of Bland & Altman (1986) based on the difference between the two measurements could be employed here rather than calculating a correlation coefficient, since the two measurements essentially measure the same thing. However, for reasons of comparability, we calculated correlations.

[10] In the second edition, however, the same trend as in the Danish data is described for the American sample; cf. Fenson et al. 2007: 101

Table 7. Estimated test-retest (Pearson's product-moment) correlation coefficient based on 1243 child-months pairs, where the successive month was also observed for the same child. LS 1: simple correlation coefficient. LS 2: partial correlation coefficient, conditioned on months. LS 3: range of single correlation coefficients within each month (M); ACSS: Subgroup of the American norming study (n=500), available ranges for the Pearson correlation coefficient calculated separately at each month are taken from Fenson et al. (1993: 68).

	LS 1	LS 2	LS 3	ACSS
Infant Part, Word Perception	.9395	.9268	.79 – .95	.80 – .90 (except for .60 in M12)
Infant Part, Word Production	.7683	.7211	.08 (M8), .17 (M9), .49 – .81 (M10 – M14)	.80 – .90 (except for .60 in M12)
Infant Part, Gestures	.9313	.8179	.73 – .86	around .80 (except for .60 in M12)
Toddler Part, Word Production	.9668	.9210	.83 (M16), .88 (M17), > .90 (M18 – M29)	> .90
Toddler Part, Complexity	.9277	.8926	missing value at M16, .15 (M17), .49 (M18), .57 (M19), > .75 (M20 – M29)	*not available*

Predictive values

In the following we present results on the diagnostic ability of an (early) CDI measurement: how well does a CDI (production) score at 16 or 22 months predict the performance six months later, that is, at 22 or 28 months[11]? We consider predictive values for both the low and high end of the score range; among the 10% of children who score lowest (highest) at 16 (22) months we estimate the probability of persisting in this group, that is, the probability of being among those 10% of children who score lowest (highest) six months later (positive predictive values, PPV)[12]. We estimate the negative predictive values (NPV), the mirror probabilities, as well: among the other 90% of children who do *not* score lowest (highest) at

[11]In principle, measurements at other time points qualify just as well; here, time points were chosen for reason of comparability (cf. Eriksson & Berglund 2005).

[12]The exact definitions of the score groups is: for the low group, below or equal to the 10% quantile at a specific month; for the high group: above or equal to the 90% quantile at a specific month.

16 (22) months we estimate the probability of persisting in *this* group, that is, the probability of being among those 90% of children who do *not* score lowest (highest) six months later (Table 8). Only children where measurements are available at both time points are included in the corresponding analysis.

Table 8. Estimated predictive values (%) for Danish (based on the Danish Longitudinal CDI study) and Swedish children; wherever available, exact numbers are added in parentheses. Results for Swedish are taken from Eriksson & Berglund (2005).

	time points	N*	PPV, low	NPV, low	PPV, high	NPV, high
Danish LS	16+22	162	50 (10/20)	95 (135/142)	41 (7/17)	93 (135/145)
Swedish	16+22		33		40	
Danish LS	22+28	129	73 (11/15)	98 (112/114)	40 (4/10)	92 (109/119)
Swedish	22+28		64		57	

* Number of children where both measurements are available

The 10% quantiles are 4, 26, and 136 words at 16, 22, and 28 months, respectively, and the 90% quantiles are 55, 384, and 692 words at these months. Not surprisingly, the negative predictive values overall are much higher than the positive predictive values. The estimates for the positive predictive values are between 33 and 73% and are based on only few children, that is, more afflicted with insecurity than their negative counterparts. It seems to be a bit easier to predict persistent low performance than persistence of high performance, especially for the later six months' time span between 22 and 28 months. Here, the positive predictive values are 73 and 64% for the Danish and Swedish low performers, respectively, as opposed to 40 and 57%, the corresponding values for the high performers.

Summary

In this section we presented a selection of characteristics of the Danish CDI which could only be evaluated based on repeated measurements[13]. Test-retest correlation coefficients of

[13]One other possible characteristic is the correlation between different sum scores with a possible delay, for

common sum scores as an indicator for their stability proved to be satisfactory and – for most scores – close to corresponding published American results. Predictive values were in line with published Swedish results, where available. However, relatively low positive predictive values indicate that one early CDI measurement at 16 or 22 months is not a good predictor of low or high performance six months later: to reliably predict a child's early lexical development from only one measurement seems a hopeless task. It might be worthwhile, though, to try basing prediction on two or more successive measurements per child (cf. Thieler 2006; Vach et al., *in prep.*).

2.4 Validity and reliability reconsidered

2.4.1 Validity

The concept of validity, that is, to put it naively, "the questionnaire's measuring what it is supposed to", is usually tied to the analysis of sum scores (cf. Streiner & Norman 1994). Within CDI data, sum scores of single parts are found to be valid (cf. section **??**), as a result, the sum score of the gesture subscales may be used to measure a child's communicative gestural ability[14]. However, this quality might not be sufficient for other, more fine-grained analyses. Studying proportional measures of vocabulary composition (based on the vocabulary checklist), especially the relative frequency of common nouns, Pine (1992) remarks that "although there is clearly evidence that maternal-report measures are useful for assessing the relative size of children's vocabularies at particular age-points, there is much less evidence to support the assumption that they are accurate in picking up variation in the actual make-up of these vocabularies" (page 76). Combining data from maternal reports with observational data, that is, adding "unreported but observed words" to words reported by the mothers, and comparing that to the words reported by the mothers alone, he finds significant differences and concludes that "the differences observed [...] reflect a tendency to under-report instances of other word-classes and thereby underestimate their relative contribution to the child's vo-

example, the correlation between the production sum score at age 16 month and the grammar-complexity score at age 30 month (cf. Hansen 2006).

[14]Interestingly, the full CDI score has never really been used to assess a child's overall communicative ability. Bornstein et al. (2004) use a combination of some scores, but mainly sum scores within single parts, describing different aspects of communication, have been the primary focus of analysis.

cabulary" (page 84). Hence, the validity of maternal report data seems to depend on word class (for further comparisons between parental reports and observational data, cf. Tardif et al. 1999).

Analysing single items of the CDI vocabulary checklists, as presented, for instance, in chapters 3 – 5, is a even more fine-grained analysis. Here, validity as "measuring what it is supposed to" would translate to "we do not care so much about an appropriate selection of words as we would want to be sure that parents answer correct on all items, or at least, answer with the same error probability on all items". That presents a much stronger assumption than "overall validity". The analysis of the *word continuity index*, the relative number of inversions per word as presented in section 2.2.5, gives a first impression since an instability of answers might reflect a higher error probability of the parents – at least, as long as it is possible to say, the child either uses the word or not, that is, excluding intermediate, fuzzy states (cf. van Geert & van Dijk 2003).

To gain more precise information, a specific study would have to be set up, comparing experimental or observational to parentally reported data on a single word level. However, the actual study design is not without complications, since, for example: the order of measurement methods probably matters, words can possibly be learned in between measurements, and items differ with respect to being suited for experiments or their likelihood to being used in a specific observational setting. Unless one employs a constant observation of a child followed by a thorough transcription of all her utterances, there is no "golden standard" method against which parental reports can be fully evaluated[15]).

In general, it would be interesting to analyse word-related factors as, for example, word class or input frequency with respect to their influence on parents' ability to remember and report words on a parental report form. This way it might be possible in the future, when analysing single items, to determine the need for adjustment for these method-related differences between items, and if needed, quantify them.

[15] A small scale study to investigate "single-item validity" has been undertaken based on data of four Danish children, two twin pairs, where both (longitudinal) spontaneous speech data and CDI parentally reported data are available. Here, the potential delay between words uttered in spontaneous speech sessions and their subsequent CDI reporting has been evaluated (cf. Bleses et al., *in press*: chapter 2.

2.4.2 Reliability

Reliability as well as validity has a naive interpretation; reliability indicates that the measurement is stable if repeated. This is a useful concept when dealing with mechanical measurements which can be repeated within a short period and are memory-free, such that the underlying measurement task is essentially the same. It is more questionable when applied to human beings filling in a questionnaire, since changes might have occurred between repeated measurements, and "true change is interpreted as measurement instability in the assessment of retest reliability" (Carmines & Zeller 1979: 39). It seems even more questionable, when data from a study which is originally intended to measure development and – sometimes rapid – change over time is also used to demonstrate reliability: does high reliability then indicate a bad measurement tool?

Following classical test theory, where a true score t is measured with random error ϵ to yield an observed score $T = t + \epsilon$, reliability is defined as $\text{var}(t)/\text{var}(T) = 1 - (\text{var}(\epsilon)/\text{var}(X))$ (e.g., Carmines & Zeller 1979). Usual investigations of reliability involve test-retest correlations (cf. section 2.3, above), internal consistency (as this presents a lower bound of reliability in the classical setting), or an overall small standard deviation (not standard error) of measurements. Again, the analyses are shaped with respect to sum scores rather than to the analysis of single items. Here, one could translate reliability (again) to stability of measurements over time: if a child once is able to say *arm*, this ability should stay. In filling in a report form, parents are employing their own, probably unconscious, reliability evaluation: some might fill in a word if they heard their child say it once, other might wait until they heard it clearly and used correctly at least three times. Similar strategies could be adopted in analysing single items in a longitudinal setting: an item is counted as produced first after parents reported it three times in a row, instead of counting as acquisition time the first instance of reporting (as in chapter 3, for instance). That would probably result in data which is more reliable but also seriously underestimating a child's performance at a given time.

Evaluating validity and reliability of single items boils essentially down to the same thing: investigate the communicative chain "child-producing, parent-listening, parent-recognising, parent-remembering, parent-reporting" in detail for each item, and determine where and why it might break. For example, whether the instability of reporting *det (this/that)* as opposed to *arm* can be explained by a break between parent-listening and parent-recognising or later, between parent-recognising and parent-remembering?

References

Altman, D.G. & Bland, J.M. (1983): Measurement in medicine: the analysis of method comparison studies. *The Statistician 32*, 307-317.

Andersen, C., Vach, W., Wehberg, S. & Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Bennett-Kastor, T. (1988): *Analyzing children's language.* Basil Blackwell.

Bland, J.M. & Altman, D.G. (1986): Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet*, i, 307-310.

Bland, J.M. & Altman, D.G. (1997): Statistics notes: Cronbach's alpha. *British Medical Journal 314*, 572.

Bleses, D., Vach, W., Wehberg, S., Faber, K. & Madsen, T.O. (in press): *Tidlig kommunikativ udvikling. Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI-forældrerapportundersøgelser af danske normalthørende og hørehæmmede børn* [Early communicative development. An instrument to describe language acquisition of Danish normal hearing and hearing impaired children based on CDI parental-report studies]. University Press of Southern Denmark, Odense.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 1): The Danish Communicative Development Inventories: validity and main developmental trends. Submitted to *Journal of Child Language.*

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 2): Building the case for a Danish delay: Early vocabulary development in Danish children in a cross-linguistic CDI-based comparison. Submitted to *Journal of Child Language.*

Bornstein, M. H., Chun-Shin, H. & Haynes, M. (2004): Specific and general language performance across early childhood: stability and gender considerations. *First Language 24*, 267-304.

Carmines, E.G. & Zeller, R.A. (1979): *Reliability and validity assessment.* Sage University Paper Series on Quantitative Applications in the Social Sciences (17). Sage Publications.

Eriksson, M. & Berglund, E. (2005): Language acquisition in Swedish children reflected by the MacArthur-Bates Communicative Development Inventories. Poster presented at the Xth International Conference for the Study of Child Language, Berlin.

Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethik, S. & Reilly, J. (1993): *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Singular Publishing Group, San Diego.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D. & Pethik, S. (1994): Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development 59(5)*, Serial no. 242.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J.S. & Bates, E. (2007): *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* Second edition. Paul H. Brookes Publishing Co.

Hansen, H. (2006): Die sprachliche Entwicklung dänischer Kinder [Language development of Danish children]. Internship report. University of Dortmund.

Lauritzen, L., Jørgensen, M., Olsen, S., Straarup, E. & Michaelsen, K. (2005): Maternal fish oil supplementation in lactation: effect on developmental outcome in breast-fed infants. *Nutrition Development 45*, 535-547.

Pine, J. (1992): How referential are "referential" children? Relationships between maternal-report and observational measures of vocabulary composition and usage. *Journal of Child Language 19*, 75-86.

Streiner, D.L. & Norman, G.R. (1995): Validity. In *Health Measurement Scales. A Practical Guide to Their Development and Use.* Second edition. Oxford University Press. Chapter 10, 144-162.

Tardif, T., Gelman, S. & Xu, F. (1999): Putting the "noun bias" in context: a comparison of English and Mandarin. *Child Development 70*, 620-635.

Thieler, A. (2006): Auswertungen zur ersten longitudinalen Studie mit der dänischen Version des MacArthur-Bates CDI [Analyses of the first longitudinal study based on the Danish version of the MacArthur-Bates CDI]. Internship report. University of Dortmund.

Vach, W., Wehberg, S. & Bleses, D. (in preparation): The growth of the vocabulary in Danish children: Results from a longitudinal CDI-based study.

van Geert, P. & van Dijk, M. (2003): Ambiguity in child language: the problem of interobserver reliability in ambiguous observation data. *First Language 23*, 259-284.

Chapter 3

Danish children's first words

Analysing longitudinal data based on monthly CDI parental reports (Paper 1)

Authors:

Sonja Wehberg[1,2], Werner Vach[2,3], Dorthe Bleses[1,3], Pia Thomsen[1,4], Thomas O. Madsen[1,3], and Hans Basbøll[1,3]

Affiliations:

[1] Institute of Language and Communication,

[2] Department of Statistics,

[3] Center for Child Language,

[4] Research Group for Language Comprehension and Language Development,

all University of Southern Denmark.

Status: Accepted by *First Language*.

ABSTRACT

Using the Danish adaptation of the MacArthur-Bates Communicative Development Inventories (CDI), first language acquisition of 183 Danish children has been studied longitudinally on a monthly basis (8-30 months). Focussing on production, we study early lexical development from the very first word until roughly 100 words are produced, dividing this period into the stages of *first-1*, *-10*, *-25*, *-50*, and *first-100* words. We analyse Danish children's first words with respect to semantic-pragmatic content, sound structure, and composition of the early lexicon based on formal linguistic categories. Comparing Danish results cross-linguistically reveals both the overall typicality of Danish children's first words as well as striking differences for some single words. (108 words)

KEYWORDS: Composition of early lexicon, cross-linguistic comparison, early lexical development, MacArthur-Bates Communicative Development Inventories (CDI)

INTRODUCTION

Even if almost all children learn to speak eventually, their first words are usually greeted with great amazement and pride by the caregivers - starting to talk seems a miracle. Linguists have always been especially fascinated by a child's first attempts to speak. Empirical research on the paths into language is extensive: some of the first systematic observations based on diary data date back to the beginning of the 17th century (e.g., on king Louis XIII's development, Ingram & Le Normand 1996); newer research methods span from experiments over structured (parental) reports to, for example, the recent exploitation of modern technology enabling dense sampling strategies of spontaneous speech.

First language acquisition of Danish has been studied before (e.g., Jespersen 1922/1964; Heger 1979; Plunkett 1986; Plunkett & Strömqvist 1992). However, so far longitudinal data has not yet been collected in a larger population which would allow for a systematic analysis of varying as well as typical features of early lexical development in Danish children. Data collection on a large scale was facilitated by the adaptation of the widely used MacArthur-Bates Communicative Development Inventories (CDI) (Fenson et al. 1993; Fenson et al. 1994) to Danish. These parental report forms address the early communicative and especially lexical development of children aged 8-15 ("Words and gestures") and 16-30 months ("Words and sentences") and were chosen as the basis for the Danish Longitudinal CDI corpus (Basbøll et al. 2002; Bleses et al. 2003; Andersen et al. 2006).

In this article our focus will be on the very early phase of lexical development, that is, the period before a child's vocabulary comprises more than one hundred words. Our main objective is to study and describe the first words produced by Danish children. Our analyses are based on the Danish Longitudinal CDI corpus since development is most appropriately studied in a longitudinal design. We will analyse some important aspects of early words - or rather of *early word targets*, that is, those adult models the children attempt to use (Vihmann 1996): their sound structure and semantic-pragmatic content as well as their form based on adult use. Regarding the latter, we will follow the approach of Bates et al. (1994) and Caselli et al. (1995) in defining the categorisation of early words.

The MacArthur-Bates CDI has been adapted to more than 30 languages so far (*www.sci.sdsu.edu/cdi/adaptations*) and results on many studies have been published (e.g. on Nordic languages: Thordardottir & Ellis Weismer (1996) - Icelandic, Eriksson & Berglund (1999) - Swedish, Laakso et. al (1999) - Finnish) which greatly supports cross-linguistic research on early lexical development; in this paper we highlight similarities and differences in first words across specific languages and thereby hope to gain insight on particularities of the Danish language and culture as well as look into whether there are certain characteristics for words to become first words. For comparison we will use results which were published on American English and Italian first words based on cross-sectional CDI data (Caselli et al. 1995), since there are no results published on longitudinal CDI data so far which would be better suited for our purposes. Having different kinds of CDI data to deal with, we come across certain comparability issues and will therefore address appropriate ways of dealing with these.

METHODS

Adaptation, data collection and participants

The adaptation of the MacArthur-Bates Communicative Development Inventories (CDI) to Danish was initiated by the *Odense Language Acquisition Project* at the University of Southern Denmark, partially in collaboration with the *Paediatric Nutrition Group*, Department of Human Nutrition at the Royal Veterinary and Agricultural University (KVL) (Lauritzen et al. 2005). To ensure cross-linguistic comparability, the Danish version aimed at being as close as possible to a translation, but taking into account linguistic and cultural differences where necessary. Details can be found in Andersen et al. (2006) and Bleses et al. (in press). In the following we will refer to the "Words and Gestures" CDI form also as the Infant part and to the "Words and Sentences" form as the Toddler part of the CDI.

Our data comes from the Danish Longitudinal CDI study which was performed at the *Center for Language Acquisition* in Odense between the years 2000 and 2002. Of approximately 1500 Danish families who had a newborn child in the first week of July 1999, 412 families were selected, stratified for the gender of the child and for the region of Denmark they lived in. After initial contact by telephone, a small questionnaire was sent to these families. Those who returned the questionnaire and agreed to participate, were then checked on exclusion criteria: the children were to be monolingually Danish living with both their parents and with no reported speech, hearing or other serious (chronic) health problems. The occurrence of *otitis* *media* followed by *tympanotomy tube* (n=5) or premature birth at a gestational age of 32 weeks or later (n=4) did not lead to exclusion. The parents of the remaining 205 families were asked to complete a CDI form in the first week of each month (rendering children precisely *x* months old), using the Infant part when the child was between 8-15 months old, and subsequently the Toddler part up to the age of 30 months. Additional information on, amongst others, the family structure, education and occupation of the parents, and the day-care situation of the child was also collected by a separate questionnaire.

Data on 183 children is available after excluding families who participated only in the first month, completed the CDI forms unsatisfactorily, or had been included incorrectly. In addition, data from one family only participating in the Infant part is excluded here. For the remaining 182 children, the distribution of *gender*, *sibling status*, *education* and *occupation* of parents is presented in Table 1. *Gender* and *sibling status* distributions agree well with corresponding numbers from the general Danish child-family population. As expected, *education* and *occupations* distributions are biased towards “higher” levels when compared to overall population numbers; they are, however, comparable to numbers from the large Danish cross-sectional CDI studies (Andersen et al. 2006).

All children were followed at least until the age of 1;08 years; 89% at least until 2;01 years and 65% completed the study at 2;06 years. All families contributed forms for at least 12 out of the possible 23 months, 73% sent in a form at least 20 times, and 47% of the families participated each month.

TABLE 1. Demographic characteristics

Gender			Sibling status			Education*			Occupation*		
	N	%		N	%		N	%		N	%
Girls	94	52%	Only child	64	35%	Basic	14	8%	Basic	24	13%
Boys	88	48%	At least 1 sibling	118	65%	Short	84	46%	Middle	70	38%
						Middle	39	21%	High	53	29%
						Long	43	24%			
						Missing/ unknown	2	1%	Missing/ unknown	35	19%

* Highest" education/occupation level of mother or father; categorisations are based upon the classification systems established by Statistics Denmark (*www.dst.dk*), which follows international guidelines (e.g., ISCO)

First words produced

In the following we will only focus on the word-checklists of the Danish CDI forms and disregard all other items in the questionnaires, which, for example, ask about gestures.

An item among the Infant part word-checklist (N=410) was in general a *first word* candidate as soon as the parents noted this word as "understood and produced" the first time. In our analyses, we take into account the *first-1 word*, the *first-5*, *-10*, *-25*, *-50*, and the *first-100 words* of the children. Since we have monthly data, we can only determine, in which month the 100th word was produced, but we cannot identify the 100th word itself. We decided to refer to all words produced prior to or in the month in which the 100th word was produced, as the *first-100 words*. Therefore, typically more than 100 items are included in the analyses of the *first-100 words* for each child. The same applies in all stages: if a child starts speaking at 0;10 and produces 3 words there, then all 3 words are included in the analyses of the very first words (*first-1 words*). If the child produces 5 new words in the next month, all 8 words are contained in the analyses of the *first-5 words* of children and so forth. Table 2 describes the relation between the nominal and actual lexicon size.

In order to obtain information as complete as possible on the first words, we use for those children not reaching a vocabulary size of 100 until 1;03 years of age also information available from the Toddler part. However, we use only information on those 410 items of the Toddler part which are also in the Infant part, and ignore all other information. This way, we render the conditions comparable for all children and all words. Nevertheless, 6 children did not reach a vocabulary size of 100 words within the study period, implying that some analyses are based on less than 182 children.

TABLE 2. Number of children (total N=182) per *first-x word* stage whose actual lexicon size is 1, 2-5, 6-10, ... words

Stage	Number of words a child actually produced								
	1	2-5	6-10	11-25	26-50	51-100	101-150	151-200	201-300
First-1 word	81	88	10	2	1				
First-5 words		46	113	18	2	3			
First-10 words			30	144	5	3			
First-25 words				16	150	12	3	1	
First-50 words				4	8	157	12	1	
First-100 words				4		6	152	18	2

Note: *x* words is the minimum in the corresponding *first-x words* stage; if a certain level was not reached, the total last size of lexicon is included instead.

Definition of vocabulary composition

Caselli et al. (1995) studied early lexical development cross-linguistically in a sample of 195 children acquiring Italian and 695 children acquiring American English, ranging from 8-16 months. They analysed data based on the Italian and American Infant part of the corresponding MacArthur-Bates Communicative Development Inventories.

In defining categories for the analysis of the composition of first words, we follow the definition of categories of Caselli et al. (1995) which are based on the original 20 thematic CDI categories. Hence we use the following narrow categorisation:

- *Sound effects and animal sounds* (*Lydeffekter og dyrelyde*) (category 1, 11 items)
- *Common nouns* (categories 2-9, 186 items):
  Animal names (*Dyrenavne*, 36), *Vehicles* (*Transportmidler*, 10), *Toys* (*Legetøj*, 8), *Food and drink* (*Mad og drikke*, 28), *Clothing* (*Tøj*, 21), *Body parts* (*Legemsdele*, 20), *Small household items* (*Små husholdningsting*, 39), *Furniture and rooms* (*Møbler og rum*, 24)
- *People* (*Mennesker*) (category 12, 30 items)
- *Games and routines* (*Leg og rutiner*) (category 13, 15 items)
- *Action words* (*Ord om handlinger*) (category 14, 53 items)
- *Descriptive words* (*Ord, der beskriver*) (category 15, 36 items)
- *Function words* (categories 16-20, 51 items):
  Words about time (*Ord om tid*, 8), *Pronouns* (*Ord, der henviser*, 11), *Question words* (*Spørge-ord*, 6), *Prepositions and locations* (*Forholdsord og lokaliteter*, 16), *Quantifiers* (*Kvantitetsord*, 10)

In analogy to Caselli et al. (1995), the CDI categories *Outside things* (*Udendørsting*) (category 10, 14 items) and *Places to go* (*Steder*) (category 11, 14 items) which constitute two separate categories in the Danish version, are not included in the narrow categorisation. However, these categories are included (as *Nominals*) in the second, broader categorisation: *Nominals* (categories 1-12, 255 items) and *Non-nominals* (categories 13-20, 155 items). In contrast to Caselli et al. (1995) we included in general the category *Words about time* (*Ord om tid*) (category 16, 8 items) in *Function words* and thus, in *Non-nominals*. This way, all available CDI items are included in at least the broad categorisation. To ensure comparability, however, this category was excluded where appropriate.

ANALYSIS METHODS AND RESULTS

The most frequent first words in Danish

Calculation of frequency and rank

Beginning with the *first-1 words*, we calculate for each word the frequency of occurrence among all 182 children: we take into account for each child (only) the CDI form at the first month, where at least one word is marked as produced by this child. We count the number of children where a specific word is marked as produced in these CDI forms and find for example that 42 out of 182 (23.1%) children said *hej (hi*) in their "first-talking" month (Table 3). We then assign ranks to the words, that is, the word with the highest frequency gets rank 1; the word with the second highest frequency gets rank 2 and so forth. In the upper part of Table 3, words with the highest frequencies among the *first-1 words* (rank 1-30) are presented together with their frequencies and ranks at later *first-words* stages. In the lower parts of Table 3, successive words are added, which are among the 30 most frequent words of the *first-5 words*, of the *first-10 words* and so forth. Ties are displayed with equal (the smallest possible) rank and sorted after the ranking at later stages.

TABLE 3. **The most frequent *first-1,5,10,25,50,-100 words*: Relative frequency *F(x)* of children (N=182) having a specific word among their *first-x words*, and Rank *R(x)* of this frequency across all words**

Word(*)	Translation	R(1)	F(1)	R(5)	F(5)	R(10)	F(10)	R(25)	F(25)	R(50)	F(50)	R(100)	F(100)
mm mm	(tastes-well sound)	1	38.5	1	71.4	1	84.1	2	95.1	4	97.8	9	97.8
hej ['hɑj̯]	hi	2	23.1	2	59.9	1	84.1	1	97.8	1	99.5	2	99.5
aarnnn	(vehicle sound)	3	21.4	3	52.2	4	68.7	10	86.8	11	91.8	12	96.2
vov ['vɔʊ̯]	(dog sound)	4	17	4	50	3	78	3	94.5	2	98.4	2	99.5
hej hej (farvel) ['hɑj̯ 'hɑj̯]	bye-bye	5	14.3	6	40.7	7	63.2	6	90.1	9	96.7	9	97.8
tak ['d̥ˢɑɡ̊]	thank-you	6	13.7	5	48.9	5	67.6	8	89	6	97.3	6	98.9
grrrr	(lion sound)	7	12.6	13	23.6	15	39	16	64.3	18	75.8	30	83.5
borte tit/titte bøh	(peekaboo/hiding game)	8	10.4	9	30.8	11	52.2	13	79.1	14	86.8	19	89.6
far ['fɑː]	father	8	10.4	11	29.1	9	61	7	89.6	6	97.3	2	99.5
mad ['mɛð]	food/mealtime	10	9.9	15	21.4	14	40.7	14	76.4	12	91.2	11	97.3
ja ['i̯ɛ]	yes	11	9.3	7	37.4	6	63.7	11	85.7	10	95.6	6	98.9
nej ['nɑi̯ʔ]	no	12	8.2	8	31.9	7	63.2	5	92.3	4	97.8	2	99.5
muh ['muː]	(cow sound)	13	7.7	12	26.4	13	48.4	12	80.8	13	87.9	13	95.6
mor ['mo̞(ː)ɐ̯]	mother	13	7.7	13	23.6	11	52.2	9	87.4	2	98.4	1	100
av ['ɑu̯ʔ]	(it-hurts sound)	15	6	10	30.2	10	57.7	4	92.9	6	97.3	6	98.9
mjav ['mi̯ɑ(ː)u̯]	(cat sound)	16	4.4	17	15.4	17	30.8	15	66.5	15	79.7	18	91.2
rap ['ʁɑb̥]	(duck sound)	17	3.8	16	17.6	16	35.7	17	60.4	17	78.6	20	89
klappe kage	(patty cake/clapping game)	17	3.8	18	11	20	19.8	25	35.7		46.7		59.9
barnets eget navn	(child's own name)	19	2.2	23	6	23	13.2		28		51.6		72.5
det ['d̥e̝]	this/that	20	1.6	20	7.1	22	15.4		31.3		42.3		52.7
bus ['b̥us]	bus	20	1.6		1.6		2.7		14.8		40.7		72.5
se ['se̝ːʔ]	(to) see	22	1.1	19	10.4	19	21.4	19	50.5	25	69.2	27	84.1
hvad ['vɛ(ð)]	what (?)	22	1.1	23	6	24	11.5		23.6		31.3		40.7
bade ['b̥ɛːðə]	(to) bathe	22	1.1	25	5.5	25	10.4	27	34.6	24	70.3	24	86.3
hund ['hunʔ]	dog	22	1.1	27	3.8	25	10.4		31.3		56.6	26	85.7
bil ['b̥iːʔl]	car	22	1.1	27	3.8		8.8	24	36.8	18	75.8	14	94
vand ['vɛnʔ]	water	22	1.1		1.1		1.1		6		28		69.8
min / mit / mine ['miːʔn 'mid̥ 'miːnə]	my	22	1.1		1.1		4.9		29.7		54.9		77.5
hest ['hesd̥]	horse	22	1.1		1.6		4.4		15.4		35.7		71.4
ko ['ɡ̊ʰo̞ːʔ]	cow	22	1.1		1.6		4.4		19.2		47.8		74.2
puste ['b̥ʰuːsd̥ə]	(to) blow	22	1.1		2.7		2.7		5.5		10.4		26.9
fugl ['fuːʔl]	bird	22	1.1		2.7		3.3		15.4		45.1		73.6
kat ['ɡ̊ʰɛd̥]	cat	22	1.1		2.7		3.3		18.1		40.1		73.1
farfar ['fɑːʔfɑ(ː)]	grandfather	22	1.1		2.7		4.9		22		39.6		59.9
shh	(be-quiet sound)		.5	21	6.6	18	22	18	56	21	73.6	24	86.3
mæh ['meː]	(sheep sound)		.5	21	6.6	20	19.8	20	48.9	27	65.4		81.3
sko ['sɡ̊o̞ːʔ]	shoe		.5	25	5.5	27	9.9	23	37.9	22	73.1	17	92.3
kiks ['ɡ̊ʰiɡ̊s]	cookie		.5	27	3.8	28	9.3	26	35.2	26	67	23	87.9
mormor ['mʊːʔ'mo̞(ː)ɐ̯]	grandmother			27	3.8		7.7		32.4	29	63.2		76.4
mælk ['melʔɡ̊]	milk				3.3	28	9.3	27	34.6		60.4	21	88.5
kykliky [ɡ̊ʰyɡ̊ʰli'ɡ̊ʰyː]	(rooster sound)		.5		3.3	28	9.3		26.4		41.2		59.3
bog ['b̥oːʔu̯]	book				1.6		5.5	21	40.1	20	75.3	15	93.4
bold ['b̥ɔlʔd̥]	ball		.5		2.2		6	22	39	16	79.1	16	92.9
bamse ['b̥ɑmsə]	teddy (bear)				2.7		6.6	27	34.6	29	63.2		81.3
is ['iːʔs]	ice		.5		1.1		6.6	30	34.1	23	70.9	21	88.5
sut ['sud̥]	pacifier		.5		2.2		7.1	30	34.1	28	63.7		82.4
op ['ɔb̥]	up				1.6		5.5	30	34.1		56.6		74.7
ble ['b̥le̝ːʔ]	diaper		.5		1.6		2.7		17		47.3	27	84.1
boller ['b̥ɔlɔ]	bread rolls				.5		2.2		16.5		48.9	27	84.1
banan [b̥ɛ'nɛːʔn]	banana				3.3		7.1		26.4		54.4	30	83.5

(*) Danish phonetic transcriptions based on adult speech (Molbæk Hansen (1990); Basbøll (2005)). [ʔ] is used for the syllabic laryngeal word prosody stød, a kind of "creaky voice".

Content analysis

Among the most frequent first words in Danish (Table 3), we find many terms most likely expressed in contexts involving social interaction: *hej (hi)* and *hej hej (bye-bye)* for greeting and parting, *ja (yes)* and *nej (no)*, *tak (thank-you)*, explicit games (*borte tit (peekaboo/hiding game)* and *klappe kage (patty cake/clapping game)*), *shh (be-quiet sound)*, *mad (food/mealtime)* and *mm mm (tastes-well sound)*, *bade (to bathe)*, *se (to see)*, *puste (to blow)*, *det (this/that)*, *hvad (what)*, *av (it-hurts sound)*, *min/mit/mine (my)*, and *op (up)*. Names for the closest caregivers and relatives (*mor (mother)*, *far (father)*, *farfar (paternal grandfather)*, and *mormor (maternal grandmother)*) and the child's own name are also early in the productive repertoire; in this respect, note that the Danish naming system for grandparents has paternal (*farmor* and *farfar*) and maternal (*mormor* and *morfar*) forms as well as neutral terms (*bedstemor* and *bedstefar*), and that all six terms are included in the Danish CDI and treated as separate items here. Terms for animals or objects close to a child's world - onomatopoetic forms as well as common nouns - form the last group. Danish children use terms referring to animals (*vov (dog sound)* and *hund (dog)*, *mjav (cat sound)* and *kat (cat)*, *muh (cow sound)* and *ko (cow)*, *grrrr (lion sound)*, *rap (duck sound)*, *mæh (sheep sound)*, *kykliky (rooster sound)*, *fugl (bird)*, and *hest (horse)*); they label common and potentially interesting objects: *bil (car)* and *aarnnn (vehicle sound)*, *bold (ball)*, *bamse (teddy bear)*, *bog (book)*, *ble (diaper)*, *bus*, *sut (pacifier)*, and *sko (shoe)*; and they know expressions for *is (ice)*, *banan (banana)*, *mælk (milk)*, *vand (water)*, *boller (bread rolls)*, and *kiks (cookie)*.

Two comments on sound structure

Looking at the phonotactic sequences of early words, or rather early word targets, we find that most of them are monosyllabic; of the 50 most frequent first words in Table 3, only 4 are decidedly polysyllabic (disregarding a final neutral vowel, which is often dropped): *mormor (maternal grandmother)*, *farfar (paternal grandfather)*, *kykliky (rooster sound)*, and *banan (banana)*, all characterised by reduplication, which is most clearly seen in the first two examples. With respect to initial consonants of the target words, we see 10 (of 42; 24%) and 9 (21%) word targets beginning with the bilabials [b] and [m], respectively, implying that these two consonants account for almost half of the items with initial consonants. All other consonants are represented with at most 5 items.

Patterns of ranks over time

The rank of the frequency of a specific word among all words is not necessarily constant from the *first-1 word* stage to the *first-100 words* stage. In Table 3, we see three different patterns: an almost stable rank over stages, losing rank, or growing in rank over stages. Examples of the first pattern (almost stable rank) are *hej (hi)* and *vov (dog sound)* – roughly speaking, (target) words showing this pattern are easily pronounced social words and widely used and known sound effects. These items are general in the sense that they are used early on by almost any child and throughout this whole early phase (roughly up to a vocabulary of a hundred words). Examples of the second pattern (decreasing rank) are *grrrr (lion sound)* and *klappe kage (patty cake/clapping game)*: more infrequently used sound effects and specific language games addressed to children follow this pattern. Like all others, these items gain in frequency over stages, but whereas they are among the most frequent words at a very early stage, they lose this position later. Possible explanations are that some children just do not acquire them or that some parents hesitate to mark them in the CDI checklists, because they regard them as too "primitive". Examples of the third pattern (increasing rank) are *bold (ball)* and *bog (book)* - words with this development pattern are mainly simple and important object and person labels. These are not among the very early words, but gain quickly in rank and frequency, ending with being essential to almost all children in the *first-100 words*. The most astonishing result is surely that *mor (mother)* and *far (father)* are not among the frequent very first early words with a high constant rank over time. Both fall into the third category of growing ranks over time: *mor (mother)* starts out with rank 13 of the *first-1 words* and gains a high rank first after the *first-25 words* stage, attaining the first rank at the last stage of about a hundred words; *far (father)* moves from rank 8 at the *first-1 words* to rank 2 at *first-100 words*.

Cross-linguistic comparison of first words

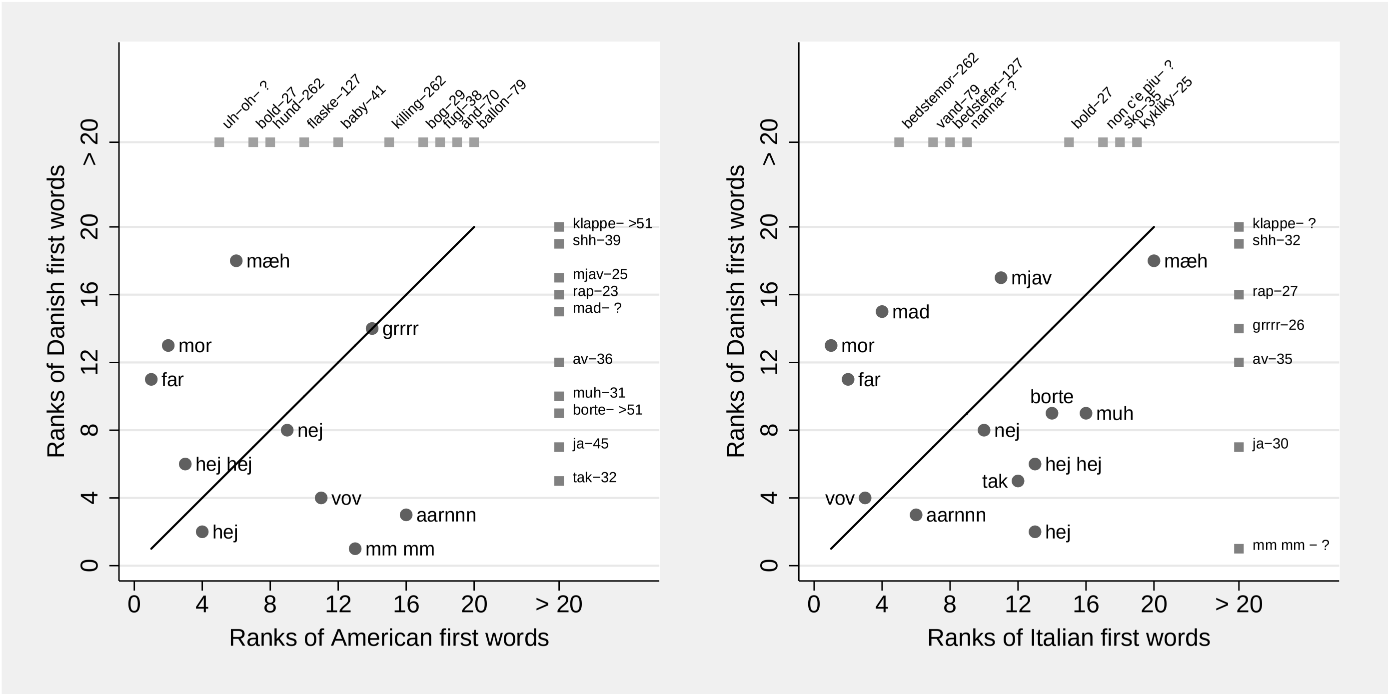
We compare our results to those published in the CDI based study of Caselli et al. (1995), who presented the "First 50 Words in Production for English versus Italian Infants" (see their Table 4, p. 186/187). Words in their lists are ranked after their frequencies as well (ranks 1-50), but there the frequency of a word was defined as the percentage of children of the total sample reported to have produced that word.

To render the ranking comparable, we imitate here the cross-sectional analysis from Caselli et al. (1995): we take all data from the Infant part (1437 CDI forms), calculate for each word the frequency of CDI forms indicating the word as marked, and rank the words after this frequency. Ties are assigned equal (the smallest possible) rank.

All Danish and American/Italian words from rank 1-20 are displayed in Figure 1. To keep the graphs "symmetric", all words out of range in the other language are displayed on the boundaries of the graphs: Words on the right boundaries have American or Italian rank > 20, respectively; words on the upper boundaries have Danish rank > 20. Words which lie roughly around the diagonal share a similar rank in both languages, for instance *nej-no* for the Danish-American as well as the Danish-Italian comparison.

Note for the left figure, that the item with American rank 5 (*uh-oh*) is missing in the Danish Infant part of the CDI and the item with Danish rank 15 (*mad (food/mealtime)*) cannot be directly identified with an item in the American CDI. For the right figure, the item with Italian rank 17 (*non c'e piu*) is not included in the Danish CDI and the items with Danish ranks 1 and 20 (*mm mm (tastes-well sound)*; *klappe kage (baking game)*) are not covered in the Italian CDI.

FIGURE 1. Comparison of ranks among Danish and American first words (Left Figure) and among Danish and Italian first words (Right Figure)



We find words which are in principle part of the (very) early vocabulary in two languages, but differ distinctly in ranks. *Mommy-mamma* and *daddy-papa* are ranked first or second in American English and Italian, which is in contrast to Danish, where names for parents emerge later (rank 13 and 11, respectively). Other examples for words which are later in Danish than in American English or Italian, are *mæh-baa baa (sheep sound)* (American English) and *mad-pappa (food/mealtime)* (Italian) (upper left quadrants). Examples for the opposite, that is early in Danish (high ranks) and later in American or Italian (lower ranks), are *mm mm-yum yum (tastes-well sound)*, *aarnnn-vroom (vehicle sound)* and *hej-ciao (hi)* (lower right quadrants in graphs).

We also find words which belong to the very early vocabulary in one language, but not in the other. There are some striking differences in ranks: *hund-dog* (ranks 262-8), *killing-kitty* (ranks 262-15) and *bedstemor-nonna (grandmother)* (ranks 262-5), *bedstefar-nonno (grandfather)* (ranks 127-8) are examples for words, which are late in Danish and early in American or Italian, respectively (upper boundaries). Words showing the opposite relation, that is early in Danish and late in American or Italian, are for instance *mad-food (mealtime)* (15-?), *ja-yes* (7-45) as well as *ja-si (yes)* (7-30) (right boundaries).

The composition of early vocabulary in Danish

As stated above, we describe the composition of the lexicon in terms of both a broad categorisation into *Nominals* and *Non-nominals* and a narrow categorisation in *Action words* and *Common nouns* etc. We calculate for each child and each *first words* stage the relative frequency of each single (or combined) category, the results of which are presented in Table 4.

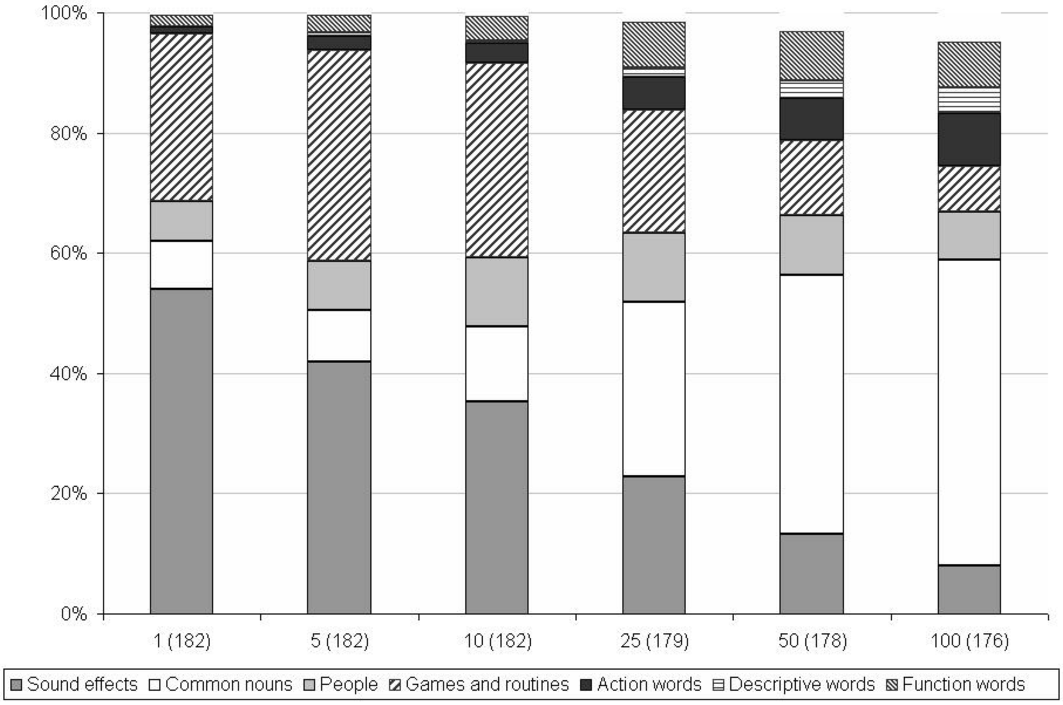
Figure 2 illustrates the development: in the (very) early stages, the majority of items in the lexicon are *Sound effects* and *Games and routines*, but their relative frequency decreases rapidly. The contribution of *Common nouns* increases continuously over stages until over 50% of the lexicon on average comes from this group. The relative frequencies of the categories *Action words*, *Descriptive words*, and *Function words* grow slowly over stages, whereas the contribution of *People* stays at a rather constant level. Using the broad categorisation, we can observe that the fraction of *Nominals* is around 60% at all stages (Table 4), implying that the sum of *Sound effects* and *Common nouns* - which are mainly *Nominals* without the *People* category - stays almost constant, too.

Table 4. Vocabulary composition at different first-x words stages: Mean values of the fraction of a category (%) within the total vocabulary of a child; corresponding 25% and 75% quantiles are presented below the means

Stages:		*1*	*5*	*10*	*25*	*50*	*100*
Number of children in stages:		182	182	182	179	178	176
Category	N*						
Nominals	255	68.9	59.1	59.7	64.8	69.4	71.6
		(33;100)	(43;75)	(50;68)	(59;71)	(64;75)	(69;75)
Non-nominals	155	31.1	40.9	40.3	35.2	30.6	28.4
		(0;67)	(25;57)	(32;50)	(29;41)	(25;36)	(25;31)
Common nouns	186	8.0	8.5	12.6	29.1	43.0	50.8
		(0;0)	(0;17)	(0;18)	(20;37)	(38;50)	(47;55)
People	30	6.6	8.2	11.4	11.4	10.0	8.1
		(0;0)	(0;17)	(6;17)	(7;14)	(7;13)	(7;10)
Sound effects	11	54.0	41.9	35.2	22.8	13.3	8.0
		(0;100)	(25;57)	(27;45)	(18;29)	(11;16)	(7;9)
Games and routines	15	27.9	35.1	32.4	20.6	12.5	7.6
		(0;50)	(20;50)	(25;40)	(17;25)	(10;15)	(7;9)
Action words	53	1.1	2.4	3.3	5.4	7.0	8.7
		(0;0)	(0;0)	(0;7)	(3;8)	(4;9)	(6;10)
Descriptive words	36	.0	.5	.5	1.5	2.8	4.3
		(0;0)	(0;0)	(0;0)	(0;3)	(0;4)	(3;6)
Function words	51	2.1	3.0	4.0	7.6	8.2	7.7
		(0;0)	(0;0)	(0;8)	(3;12)	(4;12)	(5;10)

N*: Number of words available in category (Danish CDI)

Figure 2. Vocabulary composition for the first-x words: Mean values of the fraction of a category (%) within the total vocabulary of a child; the number of children is noted behind each first words category



Note: The percentages do not add up to 100% since Places to go and Outside things contribute to the total lexicon of a child but are not included in a (narrow) category.

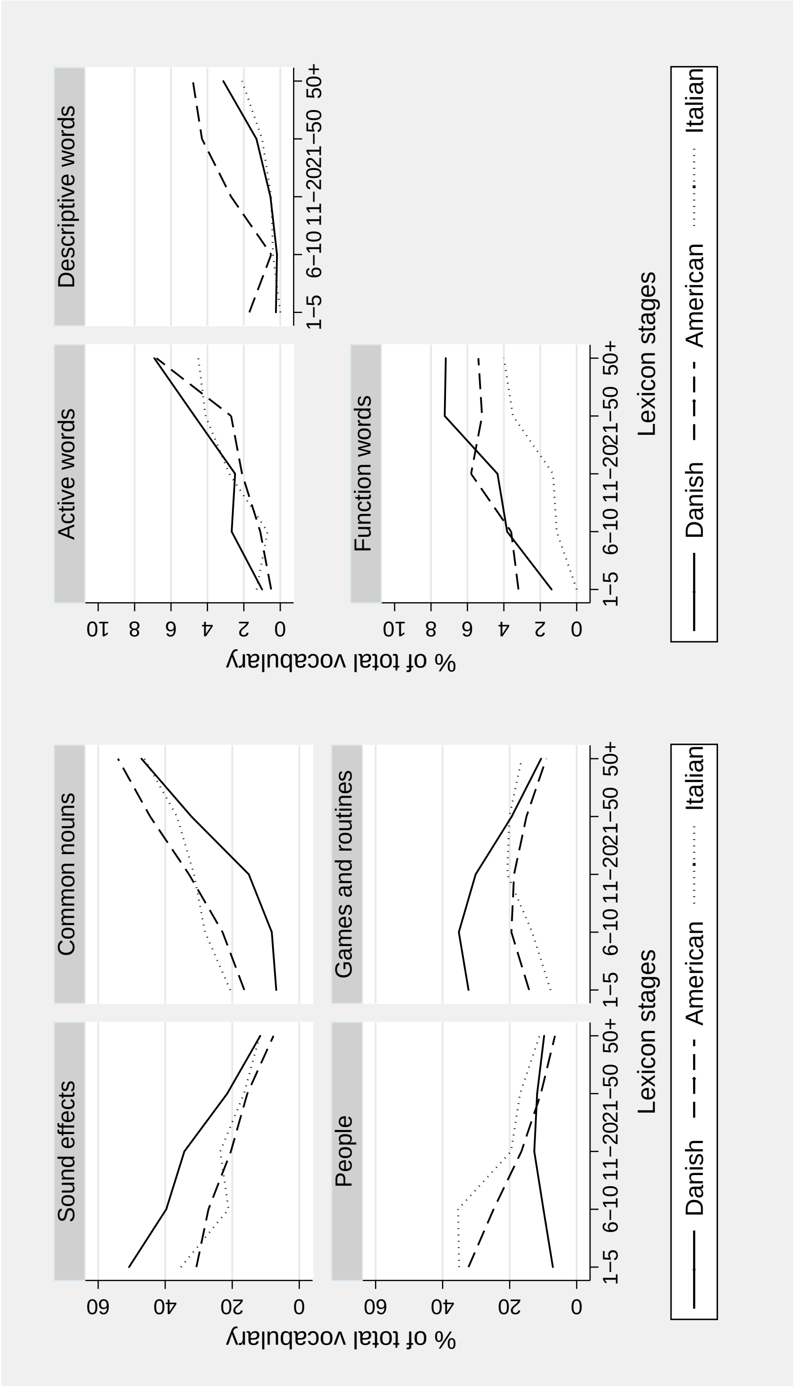
Cross-linguistic comparison of vocabulary composition

Again, we use the results of Caselli et al.'s (1995) study with Italian and American English for comparison.

The Danish results here exclude the CDI section *Time words* of the category *Function words*, and thus of *Non-nominals*, to ensure comparability to the American and Italian results from Caselli et al. (1995). Their results on lexicon composition are based on cross-sectional data, where the children were grouped according to their lexicon size: 1-5, 6-10, 11-20, 21-50, 50+. To obtain comparable frequencies, we tried to apply their approach to the CDI forms available to us. We include in our analysis all Infant forms and - in contrast to Caselli et al. (1995) - all Toddler forms as long as 100 or less words were marked, since considering just the Infant forms would have left only 3 children in the vocabulary group 50+. Note that it could happen that a child was included more than once in a vocabulary group, for example, a child producing 11 words in month 20 and 13 words in month 21 qualifies twice for the vocabulary group 11-20. In that case, a mean value for this child and vocabulary group is used in the analysis. Note also that not all children contribute in all vocabulary groups, since a child may "jump" from 9 words in month 8 to 51 in month 9.

Danish, American English, and Italian results are compared in Figure 3. The overall similarity between American and Italian is striking, particularly for the high-frequency categories. The results for all three languages converge rapidly with lexicon size, but for children with small vocabularies Danish clearly stands out. However, we must be aware of, that the very early vocabulary is small and many "essential" words (such as *mor-mommy*) are used by many children and therefore these early discrepancies are caused by only a few very early words. Figure 1 (both graphs) emphasises this tendency: for example, *mor (mother)* and *far (father)*, both included in the category *People*, are on average acquired earlier in American English and Italian than in Danish, and correspondingly, the contribution of *People* to the total vocabulary in Danish is in the early stages (much) lower than that in American English or Italian. Almost all words at the upper boundaries in both graphs of Figure 1 ("later in Danish than in American English or Italian") belong to *Common nouns*, which coincides with the finding that these are less frequent in an average Danish child's very early lexicon compared to American English or Italian children. We find the opposite relation for *Sound effects* and *Games and routines*, and correspondingly, most words at the right boundaries of the two graphs (Figure 1) can be allocated to these two categories. In general, results on single items point in the same direction, for example, *tak (thank-you)* and *ja-si (yes)* (included in *Games and routines*) are more frequent (early on) in Danish than in American English or Italian, respectively. Some items such as *av-bua (it-hurts sound)*, which are less frequent (early on) in Italian than in Danish, are classified into two different categories, into *Sound effects* in Danish and *Games and routines* in Italian. This may complicate the picture but still supports the notion that the very early differences between the languages are due to differences between very few very early single items.

FIGURE 3. Vocabulary composition (cross-sectional approach): Mean values of the fraction of high-frequency categories (%) (Left Figure) and of low-frequency categories (%) (Right Figure) within the total vocabulary of a child



SUMMARY OF RESULTS AND DISCUSSION

Cross-linguistic similarities in first words

Looking at the (adult-defined) content of first words, Dromi (1999) stated that "the contents that one-word stage speakers choose to lexicalise are remarkably similar" and "initially children acquire terms which label objects and actions that are commonly encountered in the immediate and familiar environment [...] and words which are frequently modelled to the child in repeated contexts of everyday routines (Nelson, 1985)" (p.110). In this respect, the results we presented on the content of Danish first words are "typical". Namely, the children knew names for mother and father, affirmations (*yes*) and prohibitions (*no*), they used words linked to social interaction contexts such as greeting (*hi*) and playing (*peekaboo*), objects (presumably) close to a child's world (*car* and *book*) and they talked a lot - using *Sound effects* as well as *Common nouns* - about cats, dogs and the like indicating that Danish children as well are - very early on - fascinated by fellow animates (cf. Melson 2003; Hart 2004).

With respect to the sound structure of early words in Danish we analysed the syllabic structure of adult target words and found mostly monosyllabic words, where the few polysyllabic items (e.g., *mormor* and *banana*) were characterised by reduplication. The bilabials [m] and [b] accounted for almost half of the initial consonants of the (very) first words in Danish, fitting well to results reported on American CDI data where "[w]ords beginning with b [...] make up 24% of the first 100 words" (Dale & Goodman 2005, p.73/74). Overall, our findings on sound structure are in close agreement with other research done in this area: Vihman (1996), for instance, stated that "early words are closely matched to their adult models, which are also mono- or disyllabic and typically include at most a single consonant type [...and] adult models for early words tend to conform to typical child production patterns, with more labials and dental consonants, more stops, nasals and glides than fricatives or liquids, more low and central than high or back vowels and little variation across the syllables of the word" (p.141).

Looking at the course of first words' ranks over time (from *first-1* to *first-100 words*), we established three different "ranking patterns" of Danish first words: (1) words with stable rank over stages were used (very early on) at any rate by almost any child; their adult models were mostly easily pronounced social words or frequently used sound effects. (2) Words with decreasing rank emerged "early or never" and were games, routines, or more infrequently used sound effects. And finally (3): words with increasing rank could be roughly categorised as simple and important object and person names. The CDI seemed here to capture the transition from a prelexical stage in the very beginning, when first words "are (...) essentially vocal signs, and may be compared to adult words which have very limited pragmatic range, like greetings and cries of *ouch*" (Menyuk et al. 1997, p.212) to the early *labeling phase*, involving "the completion of a two-stage shift in function from the natural use of gestures and vocal forms within the action context of a familiar routine, first to (transitional) semi-autonomous, iconic use, then to fully autonoumous (or) symbolic referential use" (Vihman 1996, p.137). Hence, the first words we were examining in this article might probably better be called *first items* since they might have been used before, during or after the transition to symbolic use and therefore with an unascertainable degree of referentiality; for reasons of simplicity and practicality however, we accepted here each CDI item as a potential first word regardless of intended meaning or use.

Regarding early *vocabulary composition*, arguments have been put forward for categorising first words in terms of adult-based formal categories (Lieven et al. 1992; Bates et al. 1994; Caselli et al. 1995) instead of classifying them with respect to child use, since the latter involves the difficult if not impossible task of revealing the communicative intentions of children with very limited verbal means (cf. Bowerman 1978; Griffiths & Atkinson 1978; Lieven et al. 1992).

Following the approach of Bates et al. (1994) and Caselli et al. (1995) who adapted the formal categorisation approach to CDI checklist data, we found that the very first words of Danish children are mainly *Sound effects*, which account on average for over 50% of the vocabulary at the *first-1 word* stage, followed by *Games and routines* with little more than 25%, and *People* and *Common nouns* - referring to objects close to a child's world - more or less equally accounting for the rest. This is in congruence with results from Caselli et al. (1995) who stated that the categories of *Sound effects*, *Proper nouns*, and *Games and routines* act as "starter sets" (p.177). Following this early phase, we saw a transition over time towards *Common nouns* at the expense of *Sound effects* and *Games and routines* which gradually lost rank and prominence over time. *Action words*, *Descriptive words*, and *Function words* first begin to pick up noticeably in frequency after the *first-50 words*, implying that at the *first-100 words Common nouns* still dominate Danish children's vocabularies with an average contribution of more than 50%. Since common nouns are widely reported to constitute the largest part of the early lexicon (cf. Bornstein et al. 2004; Dromi 1999) - preceding verbs and function words which emerge later (Caselli et al. 1995) - our findings were to be expected.

More specific comparisons between Danish results and American English and Italian results, respectively, showed up some initial differences, which were probably due to very few, very early (frequent) words: for example, the discrepancy in the category *People* might be explained by the relatively late acquisition of *mor (mother)* and *far (father)* in Danish. With increasing lexicon size, however, the results for all three languages converged rapidly, showing a strikingly similar development in spite of substantially different languages.

In light of preliminary results of the Danish CDI studies indicating that Danish children typically lag 2-3 months behind their American peers in early vocabulary comprehension and production scores (Bleses et al. 2003), the fact that we in general were able here to "re-produce" typical results on first words based on CDI studies was reassuring in two ways. First, our strategy of basing our analyses on the individual vocabulary size (within the CDI word-checklist) of the children rather than on their actual age worked well, even if we compared Danish children to potentially younger American or Italian children. This suggests that chronological age is of minor importance for those aspects of the early lexicon we are investigating here in this paper. Second, looking at the language development of Danish children within the *first-100 words* we found no support for the hypothesis that Danish children might follow a substantially different "path into the early lexicon" than their American or Italian peers. The underlying mechanisms of language acquisition seem to be fundamentally the same and therefore the potential "delay" of the Danish children seems then to be merely just that: a delay.

Cross-linguistic differences in first words

Despite the overall "typicality" of Danish first words, we also found some striking differences, for example in *bil-car-automobile*, *tak-thank-you-grazie*, *mad-food (mealtime)-pappa*, and – most prominent – *mor-mommy-mamma* and *far-daddy-papa*. Cultural differences accompanied by input frequencies and/or linguistic properties of the words, especially articulatory differences, seem plausible explanations for most of these differences, confirming that culture and language matter to first words. Assuming, however, that cultural differences play no role in acquiring names for mother and father (in these three languages), respectively, we will take a closer look at their sound structure. The Danish terms for parents have no easy consonant reduplication matching common babbling schemes as it is seen in many other languages (cf. Jakobson 1962). Danish has no alternative, simplified versions based on the nursery terms – in fact *mor* and *far* themselves are the simplified forms substituting the more traditional forms *moder* and *fader* (cf. Levin 2005). In addition, the (labiodental) fricative consonant [f] is usually seen late in acquisition, at least later than the voiced stops [b] or [d] in children acquiring English (cf. Levin 2005; Menn & Stoel-Gammon 2005), which all might account for a later acquisition of parents' names in Danish than in American English or Italian. An alternative hypothesis is that instead of Danish children being late, American and Italian children are rated as more advanced than they actually are: *mommy-mamma*, being more close to the babbling scheme [ma ma], may get "mis-interpreted" as a word by (enthusiastic) parents, rendering first words a result of social interaction. In that respect it would be interesting to explore if Danish children first producing *mor* assign more referential meaning to *mor* than their American or Italian peers do when they are first reported to address their mother and, thus, to explore if the latter occurrence should rather be attributed to the prelexical stage.

Combining the sound-related, the cultural, and the social interaction, that is, communicative aspect, two properties may be formulated which a word from any given language must fulfil in order to qualify as a potential first word: it has to be (1) close to individual or common babbling patterns and (2) it must be plausible for the addressee. The second property combines both the cultural and social interaction aspects. For example, it is only plausible for a mother to assume that her child attempts to play *peekaboo* if this game (word) exists in the respective language in the first place, and if it has been (regularly) played with the child before, that is, the context for the word to be prompted has to have been offered regularly. Within these contexts then, specific words are expected, recognised, and treated as meaningful utterances. Let us look at an example of the latter: a child utters repeatedly [ma] and her American mother recognises this as *mommy* (and not as a highly implausible *mammal*), whereas her Danish mother recognises it as meaning *mad (food/mealtime)* stimulating her child by her reaction to re-use it in meal-related situations. Instead, the child producing [ma] while sitting in the bath goes unrecognised and gets no stimulation. The parents as the usual first communication partners play the important part as (mis-)interpreters; they react with great enthusiasm to potential first words resulting in more and improved attempts of the child. Other researchers have been focussing on this "dialogic process" (Vihmann 1996, p.135) as well, Jespersen (1922/1964) for example wrote: "It is very natural that the mother who is greeted by her happy child with the sound 'mama' should take it as though the child were *calling* her 'mama', and since she frequently comes to the cradle when she hears the sound, the child himself does learn to use these syllables when he wants to call her"(p.155); cf. also Dore (1983) and Veneziano (1981). Moreover, one may even interpret the fact that many languages adopt nursery forms, and thus "build a specific infantile layer in standard vocabulary" (Jakobson 1962, p.539) accommodating children in their language learning task as well as child-adult communication in general, as a further level of this dialogic process. Again, it would be interesting to collect and analyse cross-linguistic data on parental behaviour in natural or experimental settings and examine this further.

First words: some fuzzy objects

Results on first words depend on the choice of measuring time, categorisation, study design, and last but not least on the kind of data at hand. Parentally reported CDI data have certain limitations: first, the composition and adaptation of a CDI-checklist involved choices in selecting CDI items out of all possible early utterances of children as well as grouping them into thematic categories. These systematic choices become especially visible for items most likely uttered very early on - in the Danish CDI, for example, *mm mm (tastes-well sound)* is included among *Sound effects*, *shh (be-quiet sound)* is included as a routine, and *hov (ups/being-surprised sound)* is not included in the Infant part at all - influencing the range (and composition) of possible first words available for study (cf. Vihman & McCune 1994). Second, there is no information on actual usage of a word at all in CDI data, no phonological or contextual information and no information on the assigned meaning, flexibility or frequency of use. And third, even if parents generally are reported to be valid and accurate *overall* observers of their children (cf. Fenson et al. 1994; Ring & Fenson 2000), it is also reported that different reporters report differently (Bornstein et al. 2006), and that their recall mechanisms work differently for different word classes, that is, they remembered nouns far better than verbs (Tardif et al. 1999; Pine 1992).

But the advantages are also impressive, the biggest being able to collect data on a large number of children allowing for the analysis of variation (in early lexical acquisition); as always, there is a trade-off between sample size and richness of the data. Its format especially eases studying a large(r) number of children, also in a longitudinal design. Comparing overall score results of the cross-sectional Danish CDI studies (Andersen et al. 2006) to those of the longidutinal study indicated that the (monthly) repetition in itself did not seem to have any undesired effects: the CDI works well in a longitudinal setting. Here, the longitudinal data made it possible to trace the emergence of (first) words in a child, as opposed to cross-sectional studies, where one usually has to resort to age or total vocabulary size as surrogate measures for time of (first) occurrence of words.

Our attempt of comparing our results cross-linguistically in order to set our findings in perspective presented some methodological challenges, mainly because we had to compare results on longitudinal CDI data to those on cross-sectional CDI data. To avoid comparing apples with pears, we were forced to re-analyse our data in a cross-sectional fashion, such as for example in Figure 1 compared to Table 3. In doing so we actually applied an alternative definition of first words and correspondingly, the first words, and especially their ranking position, changed considerably. Other approaches to defining first words exist in the literature, for example based upon the criterion "time (age), where at least half of the children in the given sample are reported to produce the specific word" (cf. Dale & Goodman 2005). As always, results have to be interpreted in light of the definitions employed.

In general, our hope is that work on methods which facilitate the comparability of results based on data from different sources as well as work on developing and applying methods which are especially suited to analyse longitudinal CDI data will continue and expand. In addition, our study demonstrated that analyses based on the single item level instead of the “usual” sum scores analyses can yield interesting insights, and we feel that also this line of analysis should be pursued further.

ACKNOWLEDGEMENTS

The work reported here was part of the project *Danish Children's Language Acquisition*, supported by the Carlsberg Foundation. The data collection was initiated by the *Odense Project of Language Acquisition*, supported by the Danish Research Council for the Humanities. Finally, we would like to thank everybody involved in the Danish CDI project, especially all families who participated in the labour-intensive longitudinal study. We would also like to thank all who provided helpful comments on earlier versions of this paper, especially Eva Myers Laris, Jacqueline Levin, Martin Weis Lindegaard, and Sharon Millar.

REFERENCES

Andersen, C., Vach, W., Wehberg, S., and Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1.

Basbøll, H. (2005): *The Phonology of Danish*. The Phonology of the World's Languages (ed. J. Durand). Oxford University Press.

Basbøll, H., Bleses, D., Cadierno, T., Jensen, A., Ladegaard, H.J., Madsen, T., Millar, S., Sinha, C., and Thomsen P. (2002): The Odense Language Acquisition Project. *Child Language Bulletin 22*, 11-12.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J.S., Reilly, J., and Hartung, J. (1994): Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language 21*, 85-124.

Bleses, D., Thomsen, P., Basbøll, H., Madsen, T., Vach, W., Wehberg, S., and Nielsen, C. (2003): *Midtvejsrapport: Tværsnitsundersøgelse af danske børns ordforråd i alderen 8 til 15 måneder*. Center for Sprogtilegnelse. Syddansk Universitet, Odense.

Bleses, D. Vach, W., Wehberg, S. & Faber, K. (in press): *Tidlig kommunikativ udvikling hos danske børn. Brugervejledning og teknisk manual til undersøgelser af normalthørende og hørehæmmede danske børn baseret på den danske adaption af The MacArthur-Bates Communicative Development Inventories*. University Press of Southern Denmark.

Bornstein, M. & Cote, L. with Maital S., Painter, K., Park, S., Pascual, L., Pécheux, M-G., Ruel, J., Venuti, P., and Vyt, A. (2004): Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development 75*, 1115-39.

Bornstein, M., Putnick, D., and De Houwer, A. (2006): Child vocabulary across the second year: stability and continuity for reporter comparisons and a cumulative score. *First Language 26*, 299-316.

Bowerman, M. (1978): The acquisition of word meaning: an investigation into some current conflicts. In N. Waterson & C. Snow (eds.): *The Development of Communication*. John Wiley & Sons. Chapter 4, 263-87.

Caselli, M., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., and Weir,

J. (1995): A cross-linguistic study of early lexical development. *Cognitive Development 10*, 159-99.

Dale, P. & Goodman, J. (2005): Commonality and individual differences in vocabulary growth. In M. Tomasello & D. Slobin (eds.): *Beyond Nature - Nurture: Essays in Honor of Elizabeth Bates*. Lawrence Erlbaum Associates. Chapter 2, 41-80.

Dore, J. (1983): Feeling, form, and intention in the baby's transition to language. In R. Golinkoff (ed.): *The Transition from Prelinguistic to Linguistic Communication*. Lawrence Erlbaum Associates. Chapter 10, 167-190.

Dromi, E. (1999): Early lexical development. In M. Barrett (ed.): *The Development of Language: Studies in Developmental Psychology*. Psychology Press Ltd. Chapter 5, 99-131.

Eriksson, M., & Berglund, E. (1999): Swedish early communicative development. *First Language 19*, 55-90.

Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethik, S., and Reilly, J. (1993): *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. Singular Publishing Group, San Diego.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D., and Pethik, S. (1994): *Variability in Early Communicative Development*. Monographs of the Society for Research in Child Development 59, Serial no. 242.

Griffiths, P. & Atkinson M. (1978): A "door" to verbs. In N. Waterson & C. Snow (eds.): *The Development of Communication*. John Wiley & Sons. Chapter 4, 311-9.

Hart, B. (2004): What toddlers talk about. *First Language 24*, 91-106.

Heger, S. (1979): Den fonetiske udvikling hos børn. In M. Jansen & J. Lund (eds.): *Børnenes sprog - sprogene omkring børn*. Hans Reitzels Forlag, København. Chapter 1, 13-67.

Ingram, D. & Le Normand, M-T. (1996): A diary study on the acquisition of Middle French: A preliminary report of the early language acquisition of Louis XIII. *Proceedings of the 20th Annual Boston University Conference on Language Development*, 352-63.

Jakobson, R. (1962): Why "mama" and "papa"? In R. Jakobson: *Selected writings, Vol.1: Phonological studies*. Mouton, The Hague. 538-45.

Jespersen O. (1922/1964): *Language: its Nature, Development and Origin*. George Allen & Unwin Ltd, London.

Laakso, M.-L., Poikkeus, A.-M., and Lyytinen, P. (1999). Early intentional communication as a predictor of language development in young toddlers. *First Language 19*, 207-31.

Lauritzen, L., Jørgensen, M., Olsen, S., Straarup, E., and Michaelsen, K. (2005): Maternal fish oil supplementation in lactation: effect on developmental outcome in breast-fed infants. *Nutrition Development 45*, 535-47.

Levin, J. (2005): Early lexical acquisition in 8-16-month-old Danish, American, and Italian children. Master of Arts Thesis. Department of English at the University of Southern Denmark, Odense.

Lieven, E., Pine, J., and Dresner Barnes, H. (1992): Individual differences in early vocabulary development: redefining the referential-expressive distinction. *Journal of Child Language 19*, 287-310.

Melson, G. (2003): Child companion and the human-companion animal bond. *American Behavioral Scientist 47*, 31-9.

Menn, L. & Stoel-Gammon, C. (2005): Phonological development. In J. Berko Gleason (ed.): *The Development of Language*. Allyn & Bacon, Boston. Chapter 3, 62-111.

Menyuk, P., Menn, L., and Silber, R. (1997): Early strategies for the perception and production of words and sounds. In P. Fletcher & M. Garman (eds.): *Language Acquisition: Studies in First Language Development*. Cambridge University Press. Chapter 9, 198-222.

Molbæk Hansen, H. (1990): *Dansk udtale*. Gyldendal Nordisk Forlag A/S, København.

Nelson, K. (1985): *Making sense: the acquisition of shared meaning*. Academic Press, New York.

Pine, J. (1992): How referential are "referential" children? Relationships between maternal-report and observational measures of vocabulary composition and usage. *Journal of Child Language 19*, 75-86.

Plunkett, K. (1986): Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology 27*, 64-73.

Plunkett, K. & Strömqvist, S. (1992): The acquisition of Scandinavian languages. In D. Slobin (ed.): *The crosslinguistic study of language acquisition, Vol. 3*. Lawrence Erlbaum Associates. Chapter 7, 457-556.

Ring, E.D. & Fenson, L. (2000): The correspondence betweeen parent report and child performance for receptive and expressive vocabulary beyond infancy. *First Language 20*, 141-59.

Tardif, T., Gelman, S., and Xu, F. (1999): Putting the "noun bias" in context: a comparison of English and Mandarin. *Child Development 70*, 620-35.

Thordardottir, E. & Ellis Weismer, S. (1996): Language assessment via parent report: Development of a screening instrument for Icelandic children. *First Language 16*, 265-85.

Veneziano, E. (1981): Early language and nonverbal communication: A reassessment. *Journal of Child Language 8*, 541-63.

Vihman, M. & McCune, L. (1994): When is a word a word? *Journal of Child Language 21*, 517-42.

Vihman, M. (1996): *Phonological Development: The Origins of Language in the Child*. Blackwell Publishers.

Chapter 4

Girls talk about dolls and boys about cars?

Analyses of group and individual variation in Danish children's first words (Paper 2)

Authors:

Sonja Wehberg[1,2], Werner Vach[2,3], Dorthe Bleses[1,3], Pia Thomsen[1,4], Thomas O. Madsen[1,3], and Hans Basbøll[1,3]

Affiliations:

[1] Institute of Language and Communication,

[2] Department of Statistics,

[3] Center for Child Language,

[4] Research Group for Language Comprehension and Language Development,

all University of Southern Denmark.

Status: Accepted by *First Language*.

ABSTRACT

Based on data from the Danish Longitudinal CDI study (REF companion paper) on 182 Danish children, we analyse aspects of variation in the children's first 100 words (produced). First, we demonstrate the effect of *gender* and *birth order* (number of siblings) on acquisition times of first words by identifying single words which are significantly earlier in the productive repertoire of, for instance, girls versus boys. We also investigate the effect of the same factors on the composition of the vocabulary where the definition of categories (word classes) are based on the CDI's thematic categorisation. Last, we investigate the individuality of the lexicon's composition and find time-persistent differences between children for some word classes at this early stage (117 words).

KEYWORDS: Birth order, composition of early lexicon, gender, longitudinal study, MacArthur-Bates Communicative Development Inventories (CDI), parental reports

INTRODUCTION

Variation in early lexical acquisition is huge; for instance, expressive vocabulary size for American children at age 1;6 is reported to range from 6 to 357 words (measured by the MacArthur-Bates CDI data on the American norming sample, Bates et al. 1994). Switching from the *chronological age* scale to the *vocabulary size* scale which presumably yields more homogeneous groups of children, leaves still considerable variation to be explained; for instance, the percentage of common nouns produced for children with an expressive vocabulary size of 101-200 words in the American norming study ranges from 53.6% to 98.7% (Bates et al. 1994).

To explore and explain variation, a fist possibility is to investigate the effect of (pre-defined) factors on the early lexicon such as *gender* or *birth order*. The effect of *gender* on early lexical acquisition (rate) has been widely investigated and demonstrated consistently over a wide range of studies. Studies based on CDI data or comparable vocabulary measures of a rather large sample size have found a slightly but consistently higher acquisition rate in girls than in boys for both comprehension and production (e.g., Fenson et al. 1994; Maital et al. 2000; Bornstein et al. 2004a; Szagun et al. 2006). The effect of *birth order* on lexical acquisition seems less marked and consistent. Fenson et al. (1994) reported small negative correlations between birth order and word production, indicating a slightly faster acquisition rate of first-borns. In a more recent study, Bornstein et al. (2004b) investigated the effect of *birth order* and *gender* on vocabulary competence within siblings of the same chronological age based on different methodologies - maternal report, child speech, and experimenter assessment - and reported an almost overall advantage for girls over boys (stratified for birth order), and an advantage for first-borns (within same-sex siblings) only when maternal reports were used.

The question addressed here is whether there are, apart from overall speed of lexical acquisition, any other, more qualitative differences in the (very) early lexicon of children - within their first hundred words - which can be attributed to *gender* or *birth order*. Just asking “are there any words which girls acquire faster (at a younger age) than boys?” would probably result in the answer “almost all words” since girls are generally (a bit) faster than boys, according to the results reported above. Instead, we analyse here the influence of *gender* and *sibling status* on the acquisition time as measured by *vocabulary size* of single first words. As such, we look at the individual *acquisition sequences* of children and phrase the research question as "is it possible to identify first words, which boys, for instance, tend to have earlier (as 10^{th} word, say) in their lexicon than girls (who first have it as 50^{th} word, say)?". This will be studied using statistical methods for analysing time-to-event data. Additionally, we will investigate whether there is an influence of the same child characteristics on the vocabulary composition of “word classes” such as sound effects or common nouns.

A second possibility to describe and explain variation is trying to identify groups of children sharing a common pattern of development, that is, addressing the question of individual style. Over the last three decades there has been much research on individual variation in early lexical acquisition, for example by Nelson (1973), Bates et al. (1988) or Lieven et al. (1992). In this, the variation has been formulated in terms of a referential-expressive distinction, an analytic versus holistic style, or a preference for *frozen phrases* versus *common nouns*, respectively. We mimic the approach of Lieven et al. (1992) and compare the vocabulary composition in the first 50 words of a child to that in the second 50 words in our longitudinal CDI data, that is, we look for time-persistent preferences for certain word classes within children. These may - if they exist - reflect "individual paths into language" and thus, may also be interpretable in terms of already established distinctions.

METHODS

The data

Data comes from the Danish Longitudinal CDI study (REF companion paper): 183 Danish children were followed monthly from 0;8 to 2;6 years of age, collecting data with the Infant part of the Danish CDI (cf. Andersen et al. 2006) between 0;8 and 1;3 years of age, and subsequently, with the Danish Toddler part until study end. Since one family stopped participation after the Infant part period, data on 182 children is included in the following. See REF companion paper and Andersen et al. (2006) for details on the adaptation process, data collection, participants, and compliance.

The sample consists of 88 boys (48%) and 94 girls (52%); 64 children (35%) were first-borns, whereas 118 (65%) had at least one sibling. Since we consider the binary distinction "only child" versus "child with at least one sibling", the term *sibling status* seems a more appropriate label than the presumably ordinal *birth order*. Note that since information on sibling status was collected at study start when the children themselves were 8 months old, their siblings are most likely older siblings. A table presenting the distribution of *education* and *occupation* levels of parents can be found in REF companion paper.

First words

As in REF companion paper we will focus on analysing the word-checklist parts of the CDI data. An item among the Infant part checklist (N=410) becomes part of a child's *first-100 words* as soon as the parents checked this item off (once) as "understood and produced" until or in the month, where they checked the 100th item. For children not reaching a vocabulary size of 100 until 1;03 years of age, information from the Toddler part (on Infant part items) is also used. More details are given in REF companion paper.

Definition of vocabulary composition

As in REF companion paper, single first words or items are categorised into a narrow categorisation of:

- *Sound effects and animal sounds* (*Lydeffekter og dyrelyde*) (category 1, 11 items)
- *Common nouns* (categories 2-9, 186 items):

  Animal names (*Dyrenavne*, 36), *Vehicles* (*Transportmidler*, 10), *Toys* (*Legetøj*, 8), *Food and drink* (*Mad og drikke*, 28), *Clothing* (*Tøj*, 21), *Body parts* (*Legemsdele*, 20), *Small household items* (*Smaa husholdningsting*, 39), *Furniture and rooms* (*Møbler og rum*, 24)
- *People* (*Mennesker*) (category 12, 30 items)
- *Games and routines* (*Leg og rutiner*) (category 13, 15 items)
- *Action words* (*Ord om handlinger*) (category 14, 53 items)
- *Descriptive words* (*Ord, der beskriver*) (category 15, 36 items)
- *Function words* (categories 16-20, 51 items):

  Words about time (*Ord om tid*, 8), *Pronouns* (*Ord, der henviser*, 11), *Question words* (*Spørge-ord*, 6), *Prepositions and locations* (*Forholdsord og lokaliteter*, 16), *Quantifiers* (*Kvantitetsord*, 10) and a broad categorisation into: *Nominals* (categories 1-12, 255 items) and *Non-nominals* (categories 13-20, 155 items), also including the separate categories of *Outside things* (*Udendørsting*) (category 10, 14 items) and *Places to go* (*Steder*) (category 11, 14 items).

ANALYSIS METHODS AND RESULTS

The influence of *gender* and *sibling status* on the acquisition time of single words

Estimating a Hazard Ratio

For a single word the *acquisition time* for a given child is defined in the following manner: if the word is actually produced within the *first-100 words*, we set the acquisition time to lexicon size at the end of the preceding month plus 1. Take as an example the following: a child says *dukke* (doll) for the first time in month 20, among 12 other words appearing for the first time at this time point; before month 20, the child has already produced 51 words: the acquisition time for *dukke* (doll) (and the other 12 words) is 52. For children who did not produce the word within any of their *first-x words*, the acquisition time is regarded as censored. The *censoring time* is set to 101 for children who did not produce the word in their *first-100 words*, or to their final lexicon size plus 1 if they did not reach a 100 words vocabulary within the study, thus including the information in the analysis that the event (if ever) occurred later than this censoring time.

With the times defined as such, we fit a Cox proportional hazard model for each word separately. Such a model estimates the *hazard ratio* between, for instance, girls and boys which is the ratio of the chance of producing a given word (within a small time interval under the condition it hasn't been produced before) in one group relative to the other group: a hazard ratio of 2 indicates a twice as high chance for girls compared to boys (Kalbfleisch & Prentice 2002).

To assess the overall significance of our findings we have to take the number of analyses performed into account. For *gender*, we analyse all 306 words produced by at least 10 children, following the rule of thumb that at least 10 observed events per factor must figure in the dataset. For *sibling status*, we include *gender* as an additional factor in the analysis, hence we analyse all 262 words which were produced by at least 20 children within their *first-100 words*. The Cox model allows to compute a p-value referring to a test of the null hypothesis, that the hazard ratio is equal to 1, implying no difference between the two groups for a single word. To assess the overall significance of our findings (over all words), we have to compare the observed number of words with a p-value below 5% to the expected under the null hypothesis of no difference between the groups for all words, that is, 5% of the words tested. For *gender*, we have 45 words observed with a p-value below 5% versus 15 expected (306*0.05=15.3); for *sibling status*, we observe 30 words vs. 13 expected, hence we have clear evidence for an influence of these factors on the acquisition time of some words. Tables 1 (*gender*) and 2 (*sibling status*) show all words with p-values below 5%, accompanied by the estimated hazard ratio (HR) together with its 95%-confidence interval (CI) and the p-value. We chose to present the words ranked after their p-values (instead of, for example, their estimated hazard ratios) since we are more interested in *identifying* words instead of *quantifying* a hazard ratio and want to avoid including words where a large estimated hazard ratio is based on (extreme) results of very few children; therefore we prioritise “strong evidence” higher than a large hazard ratio.

Content analysis

In the following, we use the term *word* interchangeably with *item*. We would like to remind the reader that the children may not use these words to *refer*, *label*, or *name* in the strict sense of the words.

Summarising Table 1, we find that most words which girls use earlier than boys are related to “objects to be cared for” (*dukke (doll)* and *baby*), social relations and status (e.g., *hej (hi)*, *farmor (grandmother)*, *faster (aunt)*, *min/mit/mine (my)*, babysitter's name, and the child's own name), body parts (*kind (cheek)*, *hoved (head)*, and *tunge (tongue)*) or to reading/writing/drawing (*tegne (to draw)*, *læse (to read)*, and *bog (book)*). *Ned (down)* as potential expression for intentional movement and *sove (to sleep)*, which might be used in doll-play, are also earlier in girls than boys. Some words such as *kjole (dress)*, *fin (fine)*, and *se (to see)* are likely to be linked to an imitation of caregivers who commonly use expressions like "Se din fine kjole (look at your fine dress)!".

Words which boys acquire earlier than girls are mostly names for noisy, moving objects (e.g., bus, støvsuger *(vacuum cleaner)*, *tog (train)*, *lastbil (truck)*, *bil (car)*), *aarnnn (vehicle sound)*, *brandbil (fire engine)*, *motorcykel (motorcycle)*, *flyvemaskine (aeroplane)*, *fut (train sound)*, objects they can act on (e.g., *dør (door)*, *saks (scissors)*, *hammer*, *vindue (window)*), nature and light (*lys (light)*,

TABLE 1. Influence of *gender* – girls (N=94) versus boys (N=88) – on acquisition times of single words

Word	Translation	HR	CI	P-value	Earlier among
dukke	doll	3.05	[2.02, 4.58]	$9.33 * 10^{-8}$	girls
bus	bus	0.42	[0.30, 0.60]	$1.72 * 10^{-6}$	boys
tissemand	penis/willy	0.15	[0.06, 0.37]	0.0000282	boys
støvsuger	vacuum cleaner	0.20	[0.09, 0.43]	0.0000463	boys
tog	train	0.48	[0.33, 0.69]	0.0000829	boys
lys	light	0.48	[0.34, 0.70]	0.0000885	boys
kjole	dress	7.67	[2.70, 21.84]	0.0001343	girls
lastbil	truck	0.31	[0.17, 0.57]	0.0001416	boys
baby	baby	1.80	[1.29, 2.52]	0.0005509	girls
tegne	(to) draw	2.11	[1.34, 3.35]	0.0013892	girls
bil	car	0.61	[0.45, 0.83]	0.0014655	boys
aarnnn	(vehicle sound)	0.62	[0.46, 0.84]	0.0017422	boys
dør	door	0.52	[0.33, 0.80]	0.0035271	boys
saks	scissors	0.39	[0.21, 0.75]	0.0041962	boys
brandbil	fire engine	0.26	[0.11, 0.66]	0.0043504	boys
kage(r)	cake/cookies	0.60	[0.41, 0.87]	0.0069853	boys
frø	frog	0.41	[0.22, 0.79]	0.0070291	boys
hammer	hammer	0.46	[0.26, 0.81]	0.0076802	boys
motorcykel	motorcycle	0.23	[0.08, 0.68]	0.0082783	boys
aftensmad	supper (evening meal)	0.39	[0.19, 0.79]	0.0088227	boys
fin	fine	1.86	[1.16, 2.98]	0.0100106	girls
hej	hi	1.49	[1.10, 2.02]	0.0100227	girls
læse	(to) read	2.21	[1.20, 4.08]	0.0108673	girls
træ	tree/wood	0.40	[0.19, 0.81]	0.0113902	boys
farmor	paternal grandmother	1.70	[1.13, 2.57]	0.0114752	girls
flyvemaskine	aeroplane	0.47	[0.26, 0.85]	0.0128313	boys
faster	aunt (father's sister)	3.51	[1.30, 9.45]	0.013061	girls
fut	(train sound)	0.65	[0.46, 0.91]	0.013215	boys
kød	meat	0.43	[0.22, 0.84]	0.0140731	boys
kind	cheek	2.30	[1.17, 4.53]	0.0157615	girls
gaffel	fork	0.56	[0.35, 0.90]	0.0164836	boys
bog	book	1.44	[1.06, 1.96]	0.0183824	girls
hoved	head	1.61	[1.07, 2.42]	0.0236275	girls
pingvin	penguin	0.17	[0.04, 0.80]	0.0241587	boys
min / mit / mine	my	1.46	[1.05, 2.04]	0.0247786	girls
se	(to) see	1.44	[1.04, 1.98]	0.0264605	girls
babysitters navn	babysitter's name	1.87	[1.07, 3.27]	0.0283081	girls
stjerne	star	0.35	[0.14, 0.91]	0.0313331	boys
barnets eget navn	child's own name	1.45	[1.03, 2.05]	0.0350664	girls
kaffe	coffee	0.57	[0.34, 0.96]	0.0351945	boys
lampe	lamp/light	0.51	[0.27, 0.96]	0.0353793	boys
vindue	window	0.20	[0.04, 0.92]	0.0389691	boys
tunge	tongue	2.88	[1.05, 7.92]	0.040608	girls
ned	down	1.49	[1.01, 2.21]	0.043937	girls
sove	(to) sleep	1.40	[1.00, 1.96]	0.0476136	girls

HR : Hazard Ratio

CI : Lower and upper bounds of the 95%-Confidence Interval for the Hazard Ratio

P-value : P-value of a test wrt. the null-hypothesis "HR = 1" *træ (tree/wood)*, *stjerne (star)*, and *lampe (lamp)*), some less typical animal terms which are also Danish brand names for sweets (*frø (frog)* and *pingvin (penguin)*), food-related terms (e.g., *kage (cookie)*, *aftensmad (supper)*, *kød (meat)*, *gaffel (fork)* and *kaffe (coffee)*), and finally *tissemand (penis/willy)*.

Regarding *sibling status* (Table 2), most words children with at least one sibling acquire earlier refer to the siblings directly (*søster (sister)* and *bror (brother)*) or indirectly (*bide (to bite)*, *slå (to hit)*, *av (it-hurts sound)*, *få (to get)*, and *arm* as potentially referring to a caregiver's arm and attention), where the latter group seems especially related to the competitive situation between siblings. Children with (older) siblings have words for actions, objects or routines, which usually are used (and named) around and by older children, earlier in their productive repertoire than single children (e.g., *saks (scissors)*, *hvor (where?)*, *danse (to dance)*, *tørstig (thirsty), læse (to read)*, *tak (thank-you)*, *ude (out)*, *bange (afraid)*, and *gynge (swing)*).

First-born children use mormor (mother's mother), morfar (mother's *father)*, *babysitter's name* - the last two words do not appear in Table 2; they have p-values of 0.059 and 0.061, respectively - and *arbejde (work)*) earlier than children with older siblings, which may relate to the importance of other caregivers than the parents. *baby*, *pige (girl)* and *dreng (boy)* might refer either to other children (outside the family, obviously) or to themselves. Some sound effects emerge earlier among children with no siblings (*muh (cow sound)* and *vov (dog sound)*), which could be due to "training by the parents". It could also be

TABLE 2. **Influence of *sibling status* – no siblings (N=64) versus at least 1 sibling (N=118) – on acquisition times of single words**

Word	Translation	HR	CI	P-value	Earlier among
søster	sister	29.31	[4.03, 213.04]	0.0008452	siblings
arm	arm	1.90	[1.25, 3.01]	0.0032128	siblings
bror	brother	8.33	[1.98, 35.08]	0.0038727	siblings
mormor	grandmother	0.61	[0.43, 0.85]	0.0038879	no siblings
saks	scissors	3.11	[1.44, 6.72]	0.0038952	siblings
bide	(to) bite	2.79	[1.35, 5.77]	0.005751	siblings
muh	(cow sound)	0.65	[0.47, 0.89]	0.0066231	no siblings
baby	baby	0.63	[0.45, 0.89]	.0079251	no siblings
kold	cold	0.50	[0.29, 0.86]	0.0119985	no siblings
lukke	(to) close	0.39	[0.18, 0.82]	0.0132056	no siblings
trapper	stairs	0.38	[0.17, 0.82]	0.0140648	no siblings
hvor	where (?)	2.53	[1.17, 5.47]	0.0184063	siblings
slå	(to) hit	3.57	[1.23, 10.36]	0.0190194	siblings
danse	(to) dance	1.68	[1.07, 2.63]	0.0235047	siblings
tørstig	thirsty	2.01	[1.10, 3.67]	0.0238468	siblings
læse	(to) read	2.24	[1.11, 4.50]	0.0240499	siblings
få	(to) get	2.77	[1.14, 6.74]	0.0250501	siblings
av	(it-hurts sound)	1.44	[1.04, 2.00]	0.0267495	siblings
ur	watch/clock	0.61	[0.40, 0.95]	0.0275097	no siblings
bil	car	0.70	[0.51, 0.97]	0.0293686	no siblings
gynge	swing(s)	1.54	[1.03, 2.29]	0.0333726	siblings
arbejde	work	0.61	[0.38, 0.96]	0.0340179	no siblings
pige	girl	0.64	[0.42, 0.97]	0.0342898	no siblings
vov	(dog sound)	0.72	[0.53, 0.99]	0.040678	no siblings
tak	thank-you	1.39	[1.01, 1.90]	0.0408396	siblings
ude	out	2.02	[1.03, 3.98]	0.0410736	siblings
dreng	boy	0.63	[0.40, 0.99]	0.046023	no siblings
juice	juice	0.69	[0.48, 0.99]	0.0460579	no siblings
bange	afraid	2.21	[1.01, 4.84]	0.0474441	siblings
gynge	(to) swing	1.49	[1.00, 2.21]	0.0489256	siblings

HR : Hazard Ratio

CI : Lower and upper bounds of the 95%-Confidence Interval for the Hazard Ratio

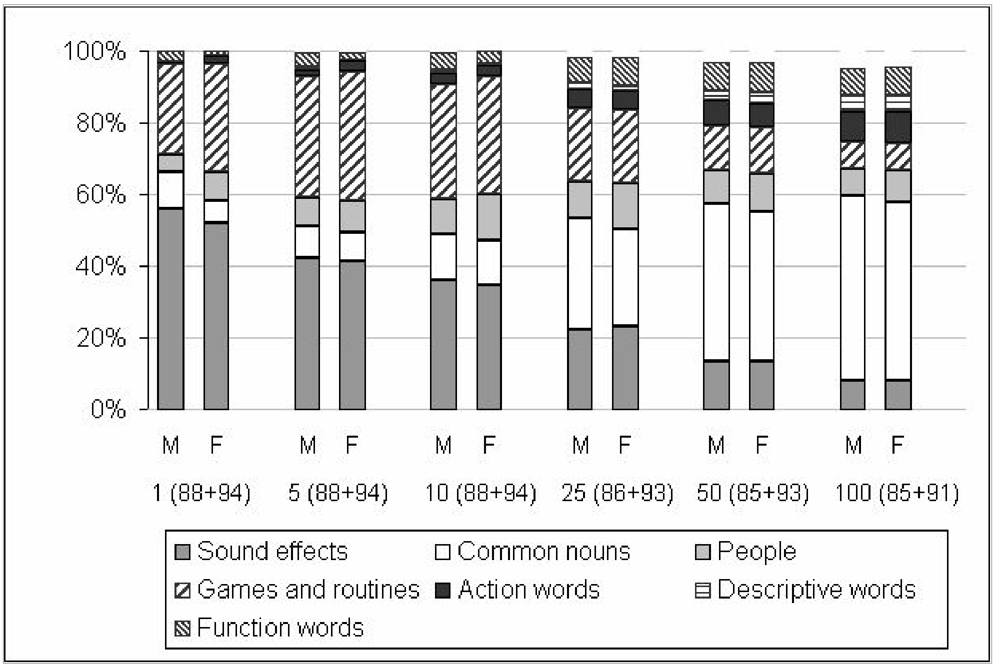
P-value : P-value of a test wrt. the null-hypothesis "HR = 1" caused by the use of the "proper" labels instead of the sounds in households with older children or it may simply reflect the relatively restricted time and attention span of parents with more than one child: these items may not be recognised, remembered or noted as words while filling out the CDI forms. Other words first-borns acquire earlier show no clear pattern; however, they again seem to relate to the increased attention and care parents tend to show their first children (e.g., *kold (cold)*, *trapper (stairs)*, *ur (clock)*, and *juice*). Words which seem to defy an ad-hoc categorisation are *bil (car)* as well as *lukke (to close)*, which are earlier in first-borns.

The influence of *gender* and *sibling status* on the composition of the lexicon

As we already in this early state could see a possible influence of child characteristics like *gender* on the acquisition time of single words, one might also consider an influence on the composition of the lexicon. Figure 1 shows the average vocabulary composition for the narrow categorisation over the stages separated by *gender*. There seems to be no substantial differences due to *gender*, and this holds true also when we look at *Nominals* and *Non-nominals*. The corresponding graphs for *sibling status* yield similar results (not shown).

These visual impressions are corroborated by a statistical analysis of the difference between groups over time: for each word class and child factor separately, we fit a binomial logistic model for the expected number of words in the word class (given the lexicon size of the child at a specific time point) with the respective child factor and the

Figure 1. Vocabulary composition for the *first-x words* by *gender* (*M* Male, *F* Female): Mean values of the fraction of a category (%) within the total vocabulary of a child; the number of boys and girls is noted behind each *first words* category



Note: The percentages do not add up to 100% since Places to go and Outside things contribute to the total lexicon of a child but are not included in a (narrow) category.

first-x words stages as categorical covariates. In order to take into account that observations within children are correlated, a robust variance estimator is used. Along with estimating the effect (odds ratio) of the specific child factor, a Wald test is used to determine whether the factor has any influence at all. There are no significant effects for the broad categories for neither *gender* nor *sibling status*, and for the narrow categories, there are only significant differences for *Common nouns*, favoured by boys, and *People*, favoured by girls and first-borns. These are, however, only associated with odds ratios less than 1.2.

Individuality in the composition of the lexicon

In the following, we correlate the fraction of words from a certain category between the first and second 50 words of a child in order to investigate whether preferences for certain categories persist over time. The *first-50 words* are defined as the first part, and all new words up to the *first-100 words* as the second part. In a first approach, we calculate Pearson's correlation coefficients for all categories (Table 3, Simple approach).

However, since the CDI measure limits the total range of words as well as that of each category, the fraction of words within the second part may suffer from a ceiling effect: if a child already in the first part has learnt many words within a category, we cannot expect the fraction of the category in the second part to be as high. For example, a child might have learnt all 11 items from the category *Sound effects* in the first part, leaving no words in the CDI category left to learn.

Consequently, there will be a tendency towards negative correlations. The effect of this is especially visible for very small categories, where almost all words were already learnt in the *first-50 words* by all children – the estimated correlation coefficient for *Sound effects*, for example, is -0.49. To correct for this ceiling effect, we estimate for each child and each category the expected fraction in the second part given the fraction in the first part and base the assessment of a persistent preference for a category over time on the difference between observed and expected fraction instead of just the observed fraction. Table 3 presents the corresponding correlation coefficients (Corrected Approach).

We find highly significant positive correlations around 0.3 for *Common nouns*, the combined group *Action*, *Descriptive*, and *Function words* as well as for *Function words* alone, and correlations ranging from 0.14 – 0.19 for the other categories, except for *Sound effects*. The correlation for *Sound effects* was -0.03, but changed to +0.11, when we removed one outlier. Overall, these findings indicate that there are moderate time-persistent preferences for certain word classes within the *first-100 words* of children.

Table 3. Pearson's correlation coefficients (CC) between first and second 50 words of children (N = 159) and corresponding p-values

	Simple Approach		Approach corrected for ceiling effect	
	CC	p-value	CC	p-value
Common nouns	-.03	.68	.32	<.0001
People	-.20	.01	.14	.08
Sound effects	-.49	<.0001	.11*	.19
Games and routines	-.17	.03	.15	.05
Action words	-.02	.83	.18	.02
Descriptive words	.03	.69	.19	.02
Function words	.07	.35	.34	<.0001
Action, Descriptive, and Function words	.06	.30	.42	<.001

* Results shown above are excluding one outlier; the correlation coefficient including the outlier is -0.03 (0.6660).

SUMMARY OF RESULTS AND DISCUSSION

The influence of gender and sibling status on first words

These two factors were chosen as exemplary child characteristics since these are commonly analysed variables in language acquisition and have been found to be of (various) impact, especially so on rate of lexical acquisition (cf. Fenson et al. 1994; Bornstein et al. 2004a, 2004b). Based on methods from time-to-event analysis, we were able to identify words which showed differences in acquisition times with respect to *gender* and *sibling status*. Post-hoc categorisations with respect to the content of the adult-target words were done to sort and structure the resulting single words: girls have, for instance, words focussing on social relations, personality, and words for "objects to be cared for" earlier in their productive repertoire than boys, whereas boys are earlier in naming for example loud, moving objects, objects they can act on and certain food-related items. Examples for word groups which first-borns use earlier than later-borns are names for caregivers other than their parents, sound effects, such as *vov (dog sound)*, as well as labels for *boy* and *girl*. Later-borns, on the other hand, know for instance names for their siblings and for objects and activities usually related to older children earlier than first-borns.

For *sibling status*, the results might be affected by a dependence of parent's form-completion behaviour on *sibling status* since parents' time and attention spans are supposedly highly influenced by the number of family members. Indeed, Bornstein et al. (2004b) found differences in vocabulary competence with respect to *birth order* in maternal report data but not in measures based on child speech or experimenter assessment. However, most of the words we identified as being earlier or later in the productive repertoire of children with or without siblings, respectively, came as no surprise at all, indicating that results on the single word level might be less afflicted by systematically different parental reporting than results on a composite level, that is, on overall production score. A probable candidate for explaining the differences we found is the (different) socialisation process for first- and later-born children involving both different environment and input (cf. Hoff-Ginsberg 1998) as well as a different competitive situation in the family triggering the need to adopt possibly different (verbal) strategies to fulfill (the same) communicative needs.

For *gender*, many of these post-hoc categories coincide with common stereotypes. A gender-related preference for toys has been reported in a variety of studies (cf. Green et al. 2004); hence the extension here, that gender-related preferences transfer to also having names for the corresponding toys earlier than the other gender, came as no surprise and suggested that the differences do not solely reflect different parental recall and form-completion strategies. Nonetheless, determining the specific cause(s) of the differences with respect to *gender* is a difficult task since it would involve disentangling the contributions of genes and socialisation (cf. Bornstein 2004a).

The ad- and post-hoc categorisation we employed served as a descriptive tool; the resulting categories are rather subjective and other researchers may come up with another categorisation of the words identified by our statistical analysis. A way of supporting the post-hoc categories is in studying whether the trends visible in these categories transfer to the "full" categories, especially in the case of those categories which defy obvious explanations - why should boys name food-related items earlier than girls, for instance? In this respect, it would also be interesting to re-address Nelson's (1973) results on the content of early referents (based on children of both sexes) that "children learn the names of things they can act on (...) as well as things that act themselves such as dogs and cars" and that "the common attribute of all of the most frequent early referents is that they have salient properties of change" (p. 31) to see if it would be possible to state (separate) preferences for girls and boys, respectively. For example, are boys in general more prone to know and name objects they can act on or that produce sound than girls?

Despite differences on (some) single words, we found nothing to support the hypothesis of structurally different vocabularies when we investigated the effect of *gender* and *sibling status* on the composition of the lexicon based on adult-defined formal categories: while children might differ with respect to *what they talked about*, we found - at this very early level - no difference in *how they talked* (according to their *gender* or *sibling status*). Since some gender differences in language style are documented for children as young as four years of age (cf. Becker Bryant 2005), our results suggest that the chosen approach of categorising words in the vocabulary is masking existing differences, which were visible on the single word level, rather than revealing them and rises the question, how gender-related language style could be defined more appropriately in terms of vocabulary composition within a CDI study.

A note of caution: in interpreting results, the potential effect of chronological age must not be disregarded, even though analyses based on the individual vocabulary size seemed more sensible to do and resulted in comparing more homogeneous groups of children. But since there is evidence that *gender* and *sibling status* groups differ with respect to their rate of acquisition (cf. Fenson et al. 1994, Bornstein et al. 2004a, 2004b), we compared potentially older boys (or later-borns) to younger girls (or first-borns). To disentangle the various impacts of chronological age, including cognitive and phonological development as well as possible corresponding changes in the environment and input, versus vocabulary size on language acquisition, presents an interesting challenge for future research.

So do (only) girls talk about dolls and boys about cars? Since in our sample all girls produced, for instance, a vehicle sound eventually, the answer is “no”. And are there boys who have *dukke (doll)* very early in their productive repertoire (acquisition sequence)? “Yes, for example, one boy says *dukke* as 16th word” – whereas 24% of the girls in our sample did not produce *dukke* amongst their *first-100 words*. However, if one poses the (longer) question whether *doll* occurs *on average* earlier in girls’ (CDI) acquisition sequences than in boys’ – or poses the corresponding question for *car* – then the answer found is “yes”. More general, we were able to show that the socialisation-related factors *gender* and *sibling status* have an effect on even the very first words children produce, even though we removed the effect of mere (higher or lower) speed of lexical acquisition by looking at the (earlier or later) time point of inclusion at a given vocabulary level. We did not find an effect of the same factors on composition of the early lexicon, indicating that differences in language style - if existent at this very early stage - should be investigated employing another categorisation approach. Hypotheses about possible causes for "why children talk earlier or later about different things" remain to be investigated.

Individuality in vocabulary composition

The real advantage of longitudinal data lies in the possibility of studying development within children. Here we studied increments of the lexicon, that is we compared the composition of the first 50 words to the composition of the newly acquired second 50 words to look for children with (persisting) preferences for certain word classes. After correcting for a potential ceiling effect in the data, we found positive correlations for all categories: correlations around 0.3 for *Common nouns*, the combined group *Action, Descriptive, and Function words* as well as *Function words* alone, and correlations ranging from 0.11 - 0.19 for the other categories (removing one outlier for *Sound effects*). Our results could be compared to those of Lieven et al. (1992), who inter alia categorised the first and second 50 words of 12 children (based on a combination of maternal reports and observed child speech) in *Common nouns*, *Frozen phrases*, and *Interactive words* etc. and found a strong correlation around 0.7 for *Frozen phrases*, followed by correlations around 0.4-0.5 for *Common nouns*, *Interactive words* and *Onomatopoetic words*, 0.2 for *Proper nouns* and no correlation for *Other words*. However, since there is no information on actual usage of words in CDI data, we had no possibility of defining a category *Frozen phrases*. Since Lieven et al. (1992) kindly provided raw data in their paper, we were able to re-calculate correlations (in their data) removing all items of the *Frozen phrases* from the individual vocabulary, which yielded smaller correlations for all remaining categories, closer to our results. However, it is important to be aware that the comparability of categories remains questionable, since (1) comparing results based on data from different sources is in general not a straightforward process (Pine et al. 1996) and (2) the different way of categorising words itself, which for example results in including other (groups of) words in Lieven et al.'s *Other words* than in our combined category *Action, Descriptive, and Function words*. Despite these limitations, results on more "comparable" categories, such as *Common nouns* and *Interactive words*, were surprisingly similar, and the discrepancy between our result on *Sound effects* compared to Lieven et al.'s (1992) result on *Onomatopoetic words* might be due to the limitation of our correction approach for very small categories.

Furthermore, one may hypothesise, that Lieven et al.'s (strong) correlation for *Frozen phrases* is linked to our findings on *Action, Descriptive, and Function words* since, in absence of contextual information in CDI data, we could resort to the probable assumption that children of this age use such words within *Frozen phrases* rather than in (productive) constructions, implying the possibility of substituting the preference for *Frozen phrases* with a preference for *Action, Descriptive, and Function words* or *Function words* alone. This, of course, remains to be studied further.

Summarising our results, we found individual time-persistent preferences for certain word categories even at this early stage in language development, demonstrating the usefulness of the CDI in this respect, even in the absence of information on *Frozen phrases*. However, it still remains to be addressed whether, and if so how, this variation in vocabulary composition might be used to identify and characterise (disjunct) groups of children with specific lexical acquisition profiles and, further, whether, and if so how - without information on the frequency of use - this variation might then be explained in terms of the widely discussed distinction of referentiality and expressiveness.

In future, methods from time-to-event analysis should definitely play a major role in analysing single CDI items in a longitudinal setting. Our application of the Cox proportional hazard model gives a first illustration, but many other research questions will require more advanced techniques. These will allow us, for example, to investigate the interrelation of acquisition times of specific items or the interrelation between age and vocabulary size as time scales, as well as they will allow us to identify subgroups of children with specific acquisition patterns, or to take the interval-censored nature of the data into account.

ACKNOWLEDGEMENTS

The work reported here was part of the project *Danish Children's Language Acquisition*, supported by the Carlsberg Foundation. The data collection was initiated by the *Odense Project of Language Acquisition*, supported by the Danish Research Council for the Humanities. Finally, we would like to thank everybody involved in the Danish CDI project, especially all families who participated in the labour-intensive longitudinal study. We would also like to thank all who provided helpful comments on earlier versions of this paper, especially Eva Myers Laris, Jacqueline Levin, Martin Weis Lindegaard, and Sharon Millar.

REFERENCES

TO BE INCLUDED: REF companion paper

Andersen, C., Vach, W., Wehberg, S., and Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1.

Bates, E., Bretherton, I., and Snyder, L. (1988): *From First Words to Grammar: Individual Differences and Dissociable Mechanisms*. Cambridge University Press.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J.S., Reilly, J., and Hartung, J. (1994): Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language 21*, 85-124.

Becker Bryant, J. (2005): Language in social contexts. In J. Berko Gleason (ed.): *The Development of Language*. Allyn & Bacon, Boston. Chapter 6, 191-229.

Bornstein, M., Hahn, C-S., and Haynes, O.M. (2004a): Specific and general language performance across early childhood: stability and gender considerations. *First Language 24*, 267-304.

Bornstein, M., Leach, D., and Haynes, O.M. (2004b): Vocabulary competence in first- and secondborn siblings of the same chronological age. *Journal of Child Language 31*, 855-73.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D., and Pethik, S. (1994): *Variability in Early Communicative Development*. Monographs of the Society for Research in Child Development 59, Serial no. 242.

Green, V., Bigler, R., and Catherwood, D. (2004): The variability and flexibility of gender-typed toy play: a close look a children's behavioral responses to counterstereotypic models. *Sex Roles 51*, 371-86.

Hoff-Ginsberg, E. (1998): The relation of birth order and socioeconomic status to children's language experience and language development. *Applied Psycholinguistics 19*, 603-29.

Kalbfleisch, J. & Prentice, R. (2002): *The statistical analysis of failure time data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Lieven, E., Pine, J., and Dresner Barnes, H. (1992): Individual differences in early vocabulary development: redefining the referential-expressive distinction. *Journal of Child Language 19*, 287-310.
Maital, S., Dromi, E., Sagi, A., and Bornstein, M. (2000): The Hebrew Communicative Development Inventory: language specific properties and cross-linguistic generalizations. *Journal of Child Language 27*, 43-67.
Nelson, K. (1973): *Structure and Strategy in Learning to Talk*. Monographs of the Society for Research in Child Development 38, Serial no. 149.
Pine, J., Lieven E., and Rowland, C. (1996): Observational and checklist measures of vocabulary composition: what do they mean? *Journal of Child Language 23*, 573-90.
Szagun, G., Steinbrink, C., Franik, M., and Stumper, B. (2006): Development of vocabulary and grammar in young German-speaking children assessed with a German language development inventory. *First Language 26*, 259-80.

Addendum to Paper 2

Addendum I. In paper 2, we fitted Cox proportional hazards models to compare the time to acquisition of a specific word between groups of children defined by *gender* or *sibling status*. In the following, we set the results of these analyses in perspective by

- describing the true structure of the data
- discussing the choice of time axis: lexicon size versus chronological age
- illustrating the use of the hazard ratio to quantify the effect
- investigating the effect of how we handled the interval-censored data

Addendum II. In analysing individual variation of first words, we corrected the correlation coefficient for a potential ceiling effect; here, we describe in detail, how we estimated the expected fraction of a certain category of words in the second 50 words (second part), taking into account that some words might already have been learnt in the first 50 words (first part).

Addendum I: Setting the analyses based on the Cox proportional hazards model into perspective

I.1 The true structure of the data

In principle, acquisition time of a specific item can be measured on two possible time axes in our study: as chronological age (month 10, say) or position in the child's acquisition sequence (word 51th, say).

Measuring time by chronological age. Children are observed monthly at the beginning of each month from 8 to at most 30 months of age, giving rise to events (= time point of first reported occurrence of a word = acquisition time) at time points: $t = 9, ..., z_i$, where $z_i \leq 30$ denotes the last observed month of child i. Information on words reported at time point 8 is left-censored (cf. Table 1), and information on words not reported until z_i is right-censored: the word could only have been acquired later, if at all. Some families failed to contribute CDI forms in some months (before resuming study participation), mostly due to vacation months or the like. Therefore, there are some (real) interval-censored observations. Obviously, the observed monthly data have a interval-censored interpretation as well: events reported in months 10, say, lie between the reporting dates of month 9 and 10. Since the same intervals are considered for all children and all words, data are grouped.

Note that the information "word A is learnt at time point 10" coincides with "word A is learnt in month 10" if one counts the first month of life as first instead of – what is usually done in calculating age – counting it as zeroth month. Hence, "word A is learnt in month 10" could be translated to "word A is learnt when the child was 9 months old", where the latter employs the usual definition of (monthly) age, that is, "after turning 9 months old" or "after having lived full 9 months". We only use the former expressions.

Measuring time by lexicon size. Lexicon size is defined as number of words reported as present in the CDI vocabulary (restricted to 410 items on the Infant vocabulary checklist). For each child, we observe if the event of interest occurred between the (individual) lexicon sizes at the

Table 1. Left-censoring: number of children N where a specific word already appears as said in month 8

Word	Translation	N
grrrr	(lion sound)	4
hej	hi	3
mm mm (lækkert)	(tastes-well sound)	3
mad	food/mealtime	2
hej hej (farvel)	bye-bye	1
ja	yes	1
mor	mother	1
nej	no	1

beginning and end of a month, that is, data are interval-censored where the intervals differ between children.

I.2 Choice of the time axis: lexicon size versus chronological age

In the paper, we have chosen lexicon size over chronological age as time axis. Intuitively, one would probably measure acquisition times by (chronological) age. However, we are interested in the effect of *gender* and *sibling status* on the acquisition times, and on the chronological age scale, girls (and to a lesser extent, also first-born children) are reported to be generally earlier and faster than boys (or later-borns) in acquiring language and specifically, in acquiring words (e.g., Fenson et al. 1994). As a consequence, the effect of *gender* and *sibling status* is presumably in favour for girls and first-born children on many words, masking the effect we are really interested in. Switching over to the lexicon size scale is a standardisation (with respect to age), rendering girls and boys (and first- and later-borns) more comparable. To gain an overview of how the two time axes correspond, we present in Table 2 information on the distribution of chronological age, at which the children participating in the Danish Longitudinal CDI study reach certain lexicon size stages (*first-x words*, cf. paper 1). Note that chronological age is given as years;months.

To illustrate the effect of a specific time axis, we compare in the following the results of our analysis based on lexicon size with respect to *gender* (cf. paper 2) to the results of a corresponding analysis based on chronological age. The comparison is based on words with a p-value below 5% for at least one of the analyses; 45 words with respect to lexicon size ("lexicon list", cf. Table 1 in paper 2) and 55 words with respect to chronological age ("age list") meet this criterium.

Before we report the results, details on model fitting will be shortly summarised. Measuring time by lexicon size (cf. paper 2), we set the acquisition time to the left interval endpoint +1, regarding the specific word acquired as first possible word in the interval. The right-censoring time was set to 101 for children who did not produce the word in their *first-100 words*, or to their final lexicon size +1, if they did not reach a 100 words vocabulary within the study. The time of acquisition of the few events which were left-censored was set to 1, collapsing the whole period before study start into a single (large) period.

Measuring time by chronological age, we set the acquisition time to the month where the word of interest was reported for the first time ($t = 8, 9, ..., z_i$), ignoring the left-censored nature of observations reported at month 8. For children who did not produce the word in their *first-100 words*, the censoring time was set to the month after when the 100th word was produced. For children who did not reach a 100 words vocabulary within the study, the censoring time was set to the last observed month +1. The few instances of (real) interval-censoring were treated as follows: if a word was reported for the first time in a month, where there was a missing CDI form the month before, acquisition time was set to the later month.

For estimation, the procedure `stcox` in Stata^{TM}was used, where ties were handled by the Peto-Breslow method. The dimension p of the covariate vector was 1 for the *gender* analyses (0=male; 1=female), and $p = 2$ for *sibling status* (0=no siblings; 1=at least 1 sibling), where we adjusted for *gender*. A model was fitted for all words where the number of observed events was at least $10 * p$, resulting in testing 306 words for *gender* and 262 words for *sibling status*. The words were then ranked after the resulting p-values (instead of, for example, their estimated hazard ratios), since we wanted to *identify* words and avoid including words where a

Table 2. Distribution of the chronological age, at which children reach the *first-x word* stages

First-x words stage	*N*	*Mean*	*SD*	*Min*	*25%–QT*	*Median*	*75%–QT*	*Max*
1	182	1;0	2.12	0;8	0;10	1;0	1;1	1;5
5	182	1;2	2.14	0;10	1;1	1;2	1;4	2;0
10	182	1;4	2.34	0;11	1;2	1;3	1;5	2;2
25	179	1;6	2.58	1;1	1;4	1;6	1;7	2;5* ($\geq 2;6$)
50	178	1;8	2.89	1;2	1;5	1;8	1;10	2;3* ($\geq 2;6$)
100	176	1;10	2.97	1;2	1;7	1;10	2;0	2;6* ($\geq 2;6$)

* The maximum is based on information of children reaching the respective *first-x words* stage. The actual maximum age for all 182 children is higher or equal than 2;6, as 2 children (excluded in Table) had 18 and 20 words, respectively, until the end of the observation period at an age of 2;6.

N : Number of children
Min : Minimum
SD : Standard deviation (based on months)
QT : Quantile
Max : Maximum

large estimated hazard ratio was based on (extreme) results of very few children.

Reporting the results, we find, not surprisingly, that a fairly large number of words appear on both lists (n=28). However, the proportion of words which are learnt (on average) earlier by boys (and correspondingly, by girls) differs considerably: 58% of the 45 words on the "lexicon list" as opposed to 20% of the 55 words on the "age list" were learnt earlier by boys. Putting it differently, all 27 words only appearing on the "age list" are learnt earlier by girls, and all 17 words only appearing on the "lexicon list" are learnt earlier by boys. A naive comparison of the estimated hazard ratios of the single words stresses this "partiality" for girl-words when using chronological age as time axis: the estimated hazard ratios on the "age list" are in general larger than on the "lexicon list"[1], indicating a more pronounced difference for words which girls learn earlier, and a less pronounced difference for words which boys learn earlier,

[1] Considering lists of words up to a p-value of 0.1, this statement holds for 58 of a total of 60 words; only *dukke (doll)* and *bog (book)* show an opposite relation.

when choosing chronological age instead of lexicon size as time axis.

The alternative approach of adjusting for lexicon size whilst keeping chronological age as time axis would be possible by including lexicon size as a time-varying covariate, for example as "lexicon size at end of previous month". In doing so, one would compare boys and girls with the same age and lexicon size, that is, fast (precocious) boys with rather slow girls; this seems problematic, to say the least.

To summarise, choosing chronological age as time axis promotes words which are learnt earlier by girls as is to be expected. Since this reflects an effect of pure speed, the results are not particularly interesting. Adjusting for lexicon size (by including it as time-varying covariate) is not likely to improve matters. The choice of lexicon size as time axis seems the only sensible alternative since this line of analysis answers a more relevant research question, namely "Is it possible to identify words, which boys tend to have earlier in their acquisition sequence than girls or vice versa?". The results, that is, the words resulting in an associated p-value below 5%, are meaningful and agree well with general knowledge about the difference between the sexes: *dukke (doll)* occurs on average earlier in girls' (CDI) acquisition sequence than in boys', and vice versa for *bil (car)*.

I.3 The use of the hazard ratio to quantify the effect

As stated above, we compare and rank words with respect to p-values which are associated to the estimated hazard ratios, summarising the available information up to the children's *first-100 words*. In choosing a Cox proportional hazards model, we implicitly assume constant hazard ratios. Obviously, this is a questionable assumption. To gain an impression on the "proportionality" of hazards, we present in Figure 1 time-varying estimates of the coefficients for *gender*, together with pointwise confidence bands, for some chosen words up to the children's *first-100 words* (cf. Table 3 for a motivation on the selection of words).

As becomes visible in Figure 1, the estimated hazard ratios may indicate some time-dependence for the eight words chosen. However, we only use the resulting averaged hazard ratios up to

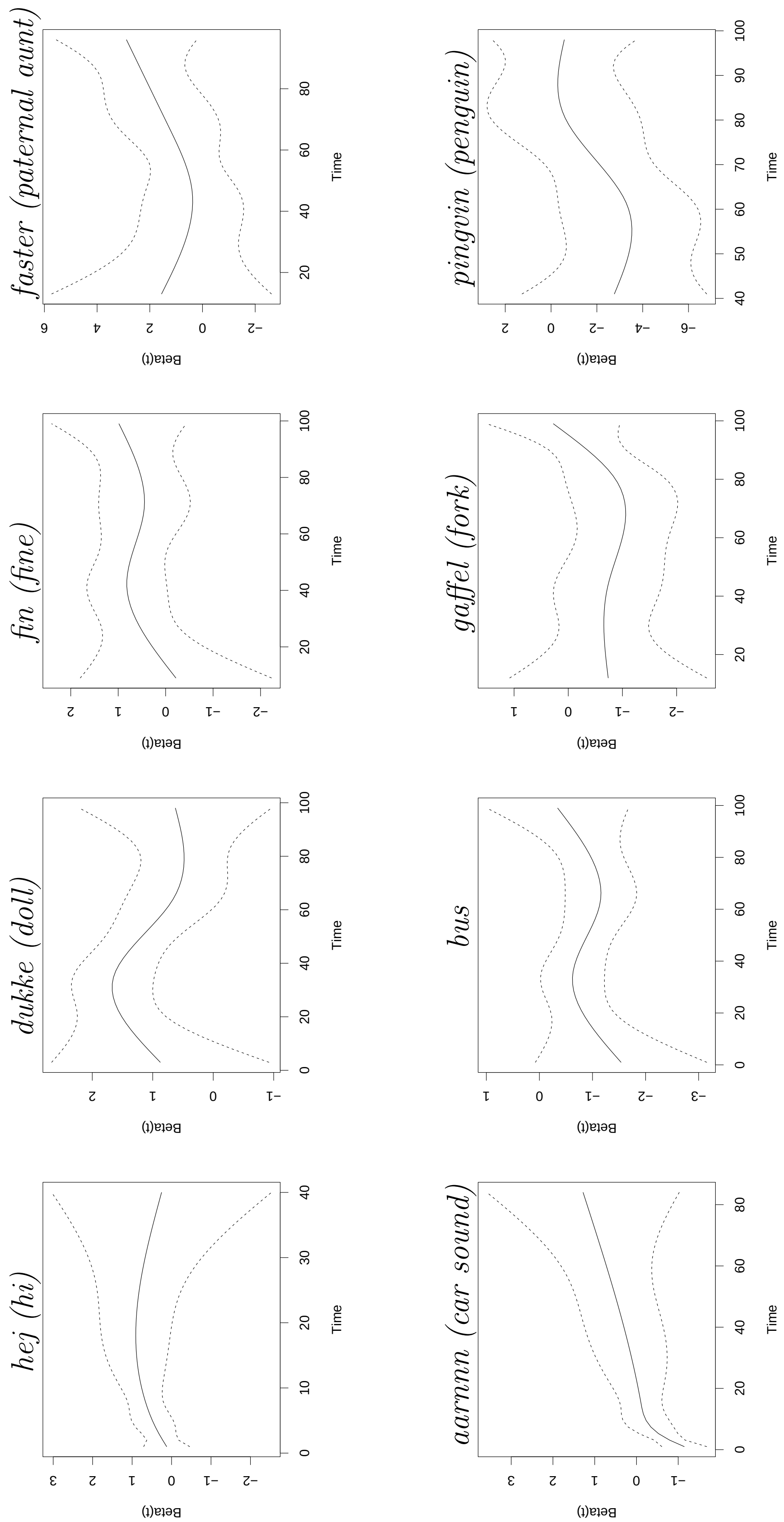


Figure 1. Time-varying estimates of coefficients up to the children's *first-100 words*. Graphs are created in R (`survival`), cf. Grambsch & Therneau 1994; Therneau & Grambsch 2001: chapter 6.

the *first-100 words* (and their associated p-values) as a sensible *criterion* for identifying and ranking the words. Then, the choice of the cut-off point matters, since it influences the number of events observed and, consequently, the power and even inclusion as we required a certain number of events to fit a model. An overview of the amount of right censoring for words with a *gender*-related p-value below 5% can be found in Table 3.

Table 3. Amount of right-censoring for *gender*-related words: Number of words (from a total of 45 words which had a p-value below 5%, cf. Table 1 in paper 2) by event rates; Information up to the *first-100 words* is considered.

N of events	% of events	N of words	% of words	Example girl-word	Example boy-word
137-182	75-100	8	18	*hej (hi)*	*aarnnn (car sound)*
91-136	50-75	10	22	*dukke (doll)*	*bus*
46-90	25-49	10	22	*fin (fine)*	*gaffel (fork)*
10-45	5-25	17	38	*faster (paternal aunt)*	*pingvin (penguin)*
Number of words where a Cox proportional hazards model was not fitted, of all 410 words:					
1-9		87			
0		17			

Our focus was to detect an *early gender*-related difference rather than a late difference, since a late difference might be due to (peculiar) children who are somewhat reluctant to learn this specific word. However, an "early" effect does not occur at the same time for all words, since some words are learnt on average very early on (for example, by a *first-100 (first-200) words*, *hej (hi)* is produced by 181 (182) children; *aarnnn (car sound)* is produced by 175 (179) children), whereas others are learnt later on (for example, by a *first-100 (first-200) words*, *faster (paternal aunt)* is produced by 23 (39) children; *pingvin (penguin)* by 12 (55) children). To capture a possible early effect even for words which are learnt relatively early in the language acquisition process resulted in the choice of the cut-off point at 100 words – under the premise that there should be a common cut-off point for all words as opposed to, say, take into account the first 100 events for each word.

To illustrate the influence of our choosing 100 as cut-off point, we present for the eight exemplary words (cf. Table 3) the estimated coefficients and hazard ratios from the Cox

Table 4. Fitting Cox proportional hazards models (in R) for eight exemplary words and the four different cut-off points 100, 200, 300, and 410.

	β	HR	SE	p-value		β	HR	SE	p-value
hej (hi)					*aarnnn (car sound)*				
100	0.436	1.55	0.156	0.0051		-0-527	0.59	0.153	0.00056
200	0.436	1.55	0.156	0.0051		-0.528	0.59	0.151	0.00048
300	0.436	1.55	0.156	0.0051		-0.513	0.598	0.151	0.00066
410	0.436	1.55	0.156	0.0051		-0.513	0.598	0.151	0.00066
dukke (doll)					*bus*				
100	1.12	3.06	0.209	8.2e-08		-0.864	0.421	0.179	1.4e-06
200	1.16	3.18	0.174	3.4e-11		-0.589	0.555	0.157	0.00017
300	1.16	3.18	0.174	3.4e-11		-0.546	0.579	0.155	0.00041
410	1.16	3.18	0.174	3.4e-11		-0.546	0.579	0.155	0.00041
fin (fine)					*gaffel (fork)*				
100	0.622	1.86	0.241	0.0098		-0.583	0.558	0.243	0.016
200	0.46	1.58	0.176	0.009		-0.494	0.61	0.162	0.0022
300	0.507	1.66	0.163	0.0019		-0.491	0.612	0.159	0.002
410	0.513	1.67	0.162	0.0015		-0.491	0.612	0.159	0.002
faster (paternal aunt)					*pingvin (penguin)*				
100	1.26	3.51	0.506	0.013		-1.75	0.174	0.775	0.024
200	0.614	1.85	0.334	0.066		-0.835	0.434	0.284	0.0032
300	0.200	1.22	0.255	0.43		-0.157	0.855	0.198	0.43
410	0.265	1.30	0.213	0.21		-0.327	0.721	0.177	0.064

β : Estimated coefficient

HR : Est. hazard ratio

SE : Standard error (of the coefficient)

proportional hazards model in Table 4. For each word, we consider 4 different cut-off points: at 100, 200, 300, and 410 words (from left to right). Not surprisingly, the estimation results of (very) early words do not change much with varying the cut-off point (e.g., the estimated coefficient for *hej (hi)* is 0.436 for all cut-off points) as opposed to later words, where there are even large discrepancies between cut-off points 300 and 410 (e.g., the estimated coefficients for *faster (paternal aunt)* decrease from 1.26 (100) over 0.614 (200) and 0.200 (300) to 0.265 (410)).

In choosing the early cut-off point of 100, we focus on the detection of a (very) early effect of *gender* whilst possibly overlooking an effect in words which are on average learnt later: there is a trade-off between being able to capture the early gender effect for early words and including (more and more events on) later words. Possibly more (later) words with a significant *gender*-related difference in acquisition time could be identified when using larger cut-off points. For example, 385 (306) words were at least learnt by 10 children until the cut-off point 200 (100) and thus, 79 additional words would have been included in the analysis based on the cut-off point 200. Of those 385 words, 72 would have resulted in a significant difference at 5% significance level (versus 385*0.05=19.2 expected)[2]. However, we risk to take "later" differences into account, when choosing a larger cut-off point. Since we wanted to avoid that, we chose the "smallest sensible" cut-off point of 100.

I.4 Short note on interval-censoring

In our approach on the lexicon size scale, we ignored the interval-censored nature of our data and simply set the acquisition time to the lexicon size at the beginning of the month, where the specific word was learnt, +1, regarding the word acquired as *first* possible word in the month.

Alternatively, one could consider the other extreme definition, regarding the specific word as acquired as *last* possible in the month. Setting the acquisition time to the *middle* of the month, is another possibility. It is also possible to take interval-censorring into account while fitting a semiparametric Cox proportional hazards model (for references, cf. the compendium for the R package `intcox`). In the following, we illustrate the implications of the various approaches by applying all four to *dukke (doll)*. Table 5 summarises the results. Since we applied the definition of acquisition time analogously in both groups, it is not surprising, that the estimated hazard ratios do not vary much between the approaches, indicating that our approach seems justifiable.

[2]Of the "original" 45 words, 37 are also amongst the 72 words; 5 words have p-values below .1, but above .05; 3 words have p-values above .1 (vindue, træ, ned).

Table 5. Fitting Cox proportional hazards models for *dukke (doll)* based on different definitions of acquisition time; Programming is done in Stata for the first, middle, and last approach. IC is calculated in R (`intcox`); confidence intervals are not included in this package.

timepoint	β	HR	SE	p-value
first	1.11	3.05	.64	9.33e-08
middle	1.26	3.51	.75	3.79e-09
last	1.17	3.22	.68	2.83e-08
IC	1.17	3.22	NA	NA

β : Estimated coefficient

HR : Est. hazard ratio

SE : Standard error (of the coefficient)

Addendum II: Correcting for a potential ceiling effect

Let N_i^1 be the number of words child i produced in the first part (*first-50 words*) and N_i^2 the number of new words child i acquired up to the *first-100 words* (second part). The corresponding inventories are defined as H_i^l, $l = 1, 2$. With $\kappa_i(W_j)$ denoting the position of word W_j ($j = 1, \ldots, 410$) in the acquisition process of child i, it follows that $W_j \in H_i^1$ if $\kappa_i(W_j) \leq N_i^1$.

Assuming that for all j, $P(\kappa_i(W_j) = k)$ does not depend on i, we can denote the probability of finding the word W_j as the kth word in the acquisition process as $p(j, k)$ for $k \geq 1$.

The probability, that the word W_j is learnt in the second part, conditioned on that is has not been learnt in the first part before, is then given by:

$$P\left(W_j \in H_i^2 \mid N_i^1, W_j \notin H_i^1, N_i^2\right)$$

$$= \frac{P\left(W_j \in H_i^2 \cap W_j \notin H_i^1 \mid N_i^1, N_i^2\right)}{P\left(W_j \notin H_i^1 \mid N_i^1, N_i^2\right)} = \frac{P\left(W_j \in H_i^2 \mid N_i^1, N_i^2\right)}{P\left(W_j \notin H_i^1 \mid N_i^1, N_i^2\right)}$$

$$= \sum_{N_i^1 < k \leq N_i^1 + N_i^2} \frac{p(j,k)}{P\left(\kappa(W_j) > N_i^1\right)} = \frac{\sum\limits_{N_i^1 < k \leq N_i^1 + N_i^2} p(j,k)}{1 - \sum_{k=1}^{N_i^1} p(j,k)}.$$

With $\mathbb{1}$ denoting an indicator function, the expected fraction of words of a category C, given the words in the first part and the size of the second part, is given by:

$$\begin{aligned}
E\,\Big(\frac{1}{N_i^2}\sum_{j\in C}\mathbb{1}(W_j\in H_i^2)\,|\,N_i^1,H_i^1,N_i^2\Big) &= E\,\Big(\frac{1}{N_i^2}\sum_{\substack{j\in C\\ j\notin H_i^1}}\mathbb{1}(W_j\in H_i^2)\,|\,N_i^1,W_j\notin H_i^1,N_i^2\Big)\\
&= \frac{1}{N_i^2}\sum_{\substack{j\in C\\ j\notin H_i^1}} P\,(W_j\in H_i^2\,|\,N_i^1,W_j\notin H_i^1,N_i^2)\\
&= \frac{1}{N_i^2}\sum_{\substack{j\in C\\ j\notin H_i^1}}\frac{\sum\limits_{N_i^1<k\le N_i^1+N_i^2} p(j,k)}{1-\sum_{k=1}^{N_i^1} p(j,k)},
\end{aligned}$$

where the first equality assumes, that the absence of specific words other than W_j in H_i^1 does not influence the probability of finding W_j in H_i^2.

We estimate $p(j,k)$ naively by

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}(W_j\in V_i(k))}{|\,V_i(k)\,|},$$

where $V_i(k)$ comprises all words, which are acquired in the same month as the kth word.

References in the Addendum

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D. & Pethik, S. (1994): Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development 59(5)*, Serial no. 242.

Grambsch, P.M. & Therneau, T.M. (1994): Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*, 515-526.

Therneau, T.M. & Grambsch, P.M. (2001): *Modeling Survival Data: Extending the Cox Model*. Springer Series: Statistics for Biology and Health. Springer.

Chapter 5

Measuring closeness of two event times

Authors:

Sonja Wehberg[1,2], Werner Vach[2,3], and Dorthe Bleses[1,3]

Affiliations:

[1] Institute of Language and Communication,

[2] Department of Statistics,

[3] Center for Child Language,

all University of Southern Denmark.

Status: To be submitted.

Abstract

This work is motivated by research in early language acquisition, where it is of interest to identify linguistic factors, which may influence whether words are acquired close together. For a pair of acquisition (event) times, we propose to base a measure of closeness on the distribution of the absolute difference $\mathfrak{D}$ between the event times, under the condition that the minimum of the two times is observed in a relevant subrange. We show how to estimate $\mathfrak{D}$ and present some candidates for absolute measures $m(\mathfrak{D})$ of closeness. We also introduce relative measures of closeness, taking the amount of chance closeness into account: we define a reference distribution $\mathfrak{D}_0$ as the distribution of the absolute difference under the assumption that the two event times are independent (conditioning correspondingly) and compare $m(\mathfrak{D})$ to $m(\mathfrak{D}_0)$. To further improve comparison, we discuss the possibility of taking covariates into account while defining and estimating $\mathfrak{D}_0$. Last, we illustrate some of the proposed closeness measures with data from the Danish Longitudinal CDI study, where acquisition times of 410 words are available for 182 children.

1 A motivating example

Even if most children seem to learn to talk easily, the question of "how they do it" is not answered as easily: research in first language acquisition is extensive. Over the years, researchers from many different fields have made their contribution to help explaining aspects of this transition to language, yet much of it remains a magical mystery. Focusing on the period where the "first words" appear, one research question still un-answered concerns the "acquisition sequence", that is, for our purpose, the sequence of words produced for the first time: for example, are there some words which are acquired close together by all (or at least by some) children? And if there are, is it then possible to identify linguistic factors, which influence whether words are acquired close together?

To investigate these questions, we analyze data of the Danish Longitudinal CDI study (cf. Andersen et al. 2006: chapter 3; Wehberg et al. 2007) which is based on the Danish adaptation of the widely-used American parental report form (Bleses et al., *in press*; Bleses et al., *submitted*), the "MacArthur-Bates Communicative Development Inventories (CDI)" (cf. Fenson et al. 1993; Fenson et al. 1994; Fenson et al. 2007). This instrument is divided into a form "Ord og gestikulation (Words and gestures)" (also called *Infant part*), which is targeted at 8-16 months old children, and a subsequent form "Ord og sætninger (Words and sentences)" (*Toddler part*) for children between 16 and 30 months of age. A central part of these CDI forms are vocabulary checklists, where parents participating in this study had to check off whether their child "understands" or "understands and says" the listed items each month anew. Items span from sound effects as *mjav (cat sound)* and language games as *klappe kage (patty cake/clapping or baking game)* over common nouns (e.g., *bed* or *elephant*), verbs and descriptive words etc. to function words as, for instance, *hvornår (when?)*. Defining the first month when an item is noted as "said" as time point of acquisition (that is, inclusion in the productive repertoire), the acquisition times of 410 common words and sound effects can be established for the 182 children participating in the study.

As a first step we want here to (1) identify those word pairs, which have a tendency to be learnt close together, and (2) quantify the extent of the closeness. Then, it becomes possible to address the second question by taking a closer look on those word pairs with a large closeness value, hoping to identify linguistic word-related features – common to several word pairs – which may explain their closeness. For example, children might learn words closer together than expected if they are connected by a semantic category as ”zoo”, or a sound-related link as between *is* and *gris* or *faster* and *pasta* (rhymes), or perhaps a formal linguistic-grammatical category (verbs, question words), or others.

2 The scope of the paper

Let $\tilde{T}_1, \tilde{T}_2$ be a pair of non-negative event times, subject to a common right-censoring time C. In principle, we want to estimate the distribution of the absolute difference $\tilde{\Delta} := |\tilde{T}_1 - \tilde{T}_2|$, that is, $\mathfrak{D}(\tilde{\Delta})$ where $\mathfrak{D}()$ denotes the distribution, and then base a measure of closeness on $\mathfrak{D}(\tilde{\Delta})$.

It is well known, that one cannot estimate "the upper tail" of the joint distribution of $\tilde{T}_1$ and $\tilde{T}_2$ from censored observations in a non-parametric way. Consequently, we cannot expect to be in general able to estimate $\mathfrak{D}(\tilde{\Delta})$. However, similar to the case of defining conditional dependence measures for bivariate failure times (cf., e.g., Oakes 1989; Fan et al. 2000), we can approach the problem by considering a conditional distribution instead, where we condition on a relevant subrange. Indeed, we show that we can estimate $\mathfrak{D}(\tilde{\Delta} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$ – at least up to a value δ^* – where $\tilde{T}_{min}$ denotes the minimum of the two event times, and t_* and t^* are chosen appropriately. Let $\mathfrak{D}$ be short for $\mathfrak{D}(\tilde{\Delta} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$.

This allows us to inspect closeness of the two event times based on absolute measures $m(\mathfrak{D})$ such as $P\,(X = 0)$ or $P\,(X \leq x)$ ($X \sim \mathfrak{D}$, x chosen appropriately), which only require an estimate of $\mathfrak{D}$ up to δ^* to plug in. However, since small values of X may happen by chance, and the amount of this chance closeness depends on the marginal distributions of $\tilde{T}_1$ and $\tilde{T}_2$, it is hard to judge $m(\mathfrak{D})$ on its own. Therefore, we show how to estimate $\mathfrak{D}_0 := \mathfrak{D}(\tilde{\Delta} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$, while assuming $\tilde{T}_1$ and $\tilde{T}_2$ independent, and compare the estimated measure $m(\hat{\mathfrak{D}})$ to $m(\hat{\mathfrak{D}}_0)$. We also consider the refinement of performing the comparison conditional on $\tilde{T}_{min}$.

Further, (some) closeness might be explained by a covariate correlated with both event times $\tilde{T}_j$ (j=1,2). To improve comparison, we discuss the possibility of taking covariates (in our example, child characteristics) into account while defining and estimating the reference distribution $\mathfrak{D}_0$.

With $\lambda(t| \,\cdot\, .)$ denoting the hazard under the condition $\cdot$, that is

$$\lambda(t \,|\, \cdot\,) = P(\tilde{T} = t, \,|\, \tilde{T} \geq t, \,\cdot\,)$$

in discrete time and

$$\lambda(t \,|\, \cdot\,) = \lim_{dt \to 0} \frac{P(\tilde{T} \in [t, t+dt] \,|\, \tilde{T} \geq t, \,\cdot\,)}{dt}$$

in continuous time, we define the following assumptions with respect to censoring:

(C1) For both $\tilde{T}_j$ (j=1,2), $\lambda(t \,|\, C \geq t) = \lambda(t) \;\forall\, t$.

(C2) For $\tilde{T}_{min}$, $\lambda(t \,|\, C \geq t) = \lambda(t) \;\forall\, t$.

(C3) For $\tilde{T}_{max}$, $\lambda(s \,|\, C \geq s, \tilde{T}_{min} = t) = \lambda(s \,|\, \tilde{T}_{min} = t) \;\forall\, s$.

(C4) $(\tilde{T}_{max} - \tilde{T}_{min}) \perp C$, given $\tilde{T}_{min} = t$. Note that $\tilde{T}_{max} - \tilde{T}_{min} = \tilde{\Delta}$.

3 Estimation of $\mathfrak{D}$

Basically, we approach this by estimating both

1. $\mathfrak{D}(\tilde{T}_{min} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$

2. $\mathfrak{D}_t := \mathfrak{D}(\tilde{\Delta} \,|\, \tilde{T}_{min} = t)$ for $t_* \leq t \leq t^*$

and combining them appropriately into an estimate for $\mathfrak{D}$, using for continuous times the relation between the densities: $f^{\,\tilde{\Delta}\,|\,t_* \leq \tilde{T}_{min} \leq t^*}(d,x) \;=\; \int f^{\,\tilde{\Delta}\,|\,\tilde{T}_{min}=t}(d,t)\,\cdot\, f^{\,\tilde{T}_{min}\,|\,t_* \leq \tilde{T}_{min} \leq t^*}(t,x)\,dt$; and for discrete times the corresponding relation:
$P(\tilde{\Delta} = d \,|\, t_* \leq \tilde{T}_{min} \leq t^*) \;=\; \sum_{t_* \leq t \leq t^*} P(\tilde{\Delta} = d \,|\, \tilde{T}_{min} = t) \,\cdot\, P(\tilde{T}_{min} = t \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$.

Note that the lower border t_* might be negligible in most applications, but is convenient in our example.

3.1 Estimation of $\mathfrak{D}(\tilde{T}_{min} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$

Assuming (C2), Kaplan-Meier provides us with an estimate of the distribution function of $\tilde{T}_{min}$, which is discrete (at least) up to the observed maximum $m^*_{1,2} := \max_i (T^i_{min})$ with $i = 1, ..., n$ observations, where $T^i_{min} := \min_i(\tilde{T}_{min}, C)$. Mass $\tilde{p}_k$ is assigned to all K observed ordered event times t_k, $k = 1, ..., K$. Some mass $\tilde{p}_{K+1}$ is left un-specified if the largest observation is right-censored.

If t_* and t^* are chosen as $t_* \leq t^* \leq m^*_{1,2}$, we obtain a discrete estimate of $\mathfrak{D}(\tilde{T}_{min} \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$, where at all observed time points $t_* \leq t_k \leq t^*$ the corresponding mass p_k is given by $p_k := \dfrac{\tilde{p}_k}{\sum_{t_* \leq t_k \leq t^*} \tilde{p}_k}$.

3.2 Estimation of $\mathfrak{D}_t$

Observations of the absolute difference $\tilde{\Delta}$ are only possible if $\tilde{T}_{min}$ is actually observed. However, even when $\tilde{T}_{min}$ is observed, $\tilde{\Delta}$ might possibly be right-censored at $C^{\Delta} := C - \tilde{T}_{min}$ if $\tilde{T}_{max} \geq C$.

Let Δ denote the observation $\min(\tilde{\Delta}, C^{\Delta})$. Then assumption (C3) allows us to obtain estimates of $\mathfrak{D}_t$ from all subjects with $\tilde{T}_{min} = t$, and assumption (C4) allows us to estimate $\mathfrak{D}_t$ from the uncensored observations with $T_{min} = t$.

Taking the approach of estimating all $\mathfrak{D}_t$ ($t_* \leq t \leq t^*$) simultaneously, one could assume a parametric model and obtain continuous estimates for $\mathfrak{D}_t$. Here, we focus on discrete estimates – in the case of both discrete and continuous event times – arising from semiparametric models such as the Cox proportional hazards model with

$$\lambda(d \mid t) = \lambda_{\text{baseline}}(d) \cdot g_\beta(t),$$

where the conditional survival function can be estimated according to Kalbfleisch & Prentice (2002: 115). For example, the function $g_\beta(t)$ could be defined as $\exp(\beta \cdot t)$. We obtain a discrete estimate for $\mathfrak{D}_{t_k}$ (at least up to the observed maximum $\delta^* := \max_i \Delta_i$), assigning probability mass $\tilde{q}_d(k)$ to each d in the set of observed values $d_1, ..., d_L$ of $\tilde{\Delta}$ (note that this set is independent of t_k).

Combining the estimates, $\mathfrak{D}$ can be approximated by a discrete estimate putting mass

$$q_d := \sum_{t_* \leq t_k \leq t^*} p_k \cdot \tilde{q}_d(k),$$

on any d in the set $\{d_1, ..., d_L\}$.

The reader may ask why we focus on joint modeling instead of considering separate Kaplan-Meier estimates for $\mathfrak{D}_{t_k}$ for each t_k. In doing so, we would get a set of $\tilde{\Delta}$-values $d_1^k, ..., d_{L_k}^k$ which depends on t_k, and corresponding probabilities masses $\tilde{q}_d(k)$. However, some mass is left un-specified if the observed maximum d_k^* of the absolute differences given t_k is censored. Therefore, the final estimate for $\mathfrak{D}$ is only "available" up to $\min(d_k^*)$, which might be close to zero. To approach useful final estimates, we need to "borrow from neighboring time points" (in t) in estimating $\mathfrak{D}_t$, and joint modeling is one way to achieve this, especially ensuring that the final estimate for $\mathfrak{D}$ is available up to δ^*. In addition, since we assume in general only a slight dependence of $\mathfrak{D}_t$ on t, it seems sufficient to allow changes in "location" (but not in "shape") as done in applying the Cox proportional hazards model.

3.3 Absolute measures

In our example, we want to compare closeness across word pairs to identify word pairs, which are learnt closer together than others. Based on $\mathfrak{D}$ alone, we would expect a higher probability on zero or small differences for close word pairs as opposed to distant word pairs. Note that our aim is different from that of measuring the agreement between two event times, although analysis of agreement should be based on the distribution of the difference as well (Bland & Altman 1986). However, in analysing agreement, one typically intends to show that a large discrepancy between two variables is unlikely, that is, one is interested in the upper tail of the distribution of the absolute difference. Here, our aim is to capture word pairs which are learnt closely together by at least some children, so our focus is on the lower tail of the distribution.

Table 1 presents some potential candidates for absolute measures $m(F)$ based on a general distribution function F (imagine $\mathfrak{D}$). Since all moment-related functionals in Table 1 cover the "classical agreement" rather than our aim and are hard to estimate from censored observations without requiring parametric modeling, these will not be considered further. The second group of "non-parametric" functionals has the advantages of (1) being simple and easy to interpret, and (2) being (in principle) evaluable with censored data (as can m_{Hx}). However, for m_x and m_{Hx}, there is choice involved in as "how far away from zero" differences are considered to be "small" to a relevant degree. Regarding the median, this choice (as well as the expectation, say) might not be sensitive enough to capture closeness due to relevant subgroups.

In the following, our focus will be on m_x, covering also the special case m_0. Such a measure can be easily computed from an estimate for $\mathfrak{D}$, provided that $x \leq \delta^*$ (when the largest observation is censored).

Table 1. Some candidates for absolute measures of closeness $m(F)$

	Candidate	Short name
Moment-related functionals:	Expectation	m_E
	Coefficient of variation	$m_{VC} := \sqrt{Var} \, / \, E$
"Non-parametric" functionals:	Tie index	$m_0 := P_F \, (\tilde{\Delta} = 0)$
	x-Prob	$m_x := P_F \, (\tilde{\Delta} \leq x)$
	Median	m_M
	IQ-range	m_{IQ}
	x-integrated hazard (rate)	$m_{Hx} := \int_0^x \lambda(s) \, ds$

4 Correcting for closeness by chance

Basing judgment of closeness on an absolute measure could be difficult, especially when considering different pairs, since small values of $\tilde{\Delta}$ may happen by chance. To put it differently, it is hard to judge (in the discrete case) whether $P(\tilde{\Delta} = 0 \,|\, t_* \leq \tilde{T}_{min} \leq t^*)$ is large or not, since the amount of chance closeness depends on the marginal distributions of $\tilde{T}_1$ and $\tilde{T}_2$:

- If $\tilde{T}_1$ and $\tilde{T}_2$ differ (substantially) by location, closeness by chance may be very small.
- Otherwise, if $\tilde{T}_1$ and $\tilde{T}_2$ have similar location and small variation, closeness by chance is large; at any rate, larger than if $\tilde{T}_1$ and $\tilde{T}_2$ have similar location and large variation.

To work around this problem, the basic idea is to compare $m(\mathfrak{D})$ to $m(\mathfrak{D}_0)$, where $\mathfrak{D}_0$ is a reference distribution which reflects the situation that closeness occurs by chance only. A first choice would be to consider $\mathfrak{D}_0$ as the distribution of $\tilde{\Delta}$ under the assumption of independence of $\tilde{T}_1$ and $\tilde{T}_2$:

$$\mathfrak{D}_0 := \mathfrak{D}_{\tilde{T}_1 \perp \tilde{T}_2}(\tilde{\Delta} \,|\, t_* \leq \tilde{T}_{min} \leq t^*).$$

4.1 Estimation of $\mathfrak{D}_0$

Assuming (C1), we propose to estimate $\mathfrak{D}_0$ by estimating first $\mathfrak{D}(\tilde{T}_j)$ (j=1,2) separately by Kaplan-Meier, then building the cross-product, and summing up appropriately. A short description follows.

Estimating the distribution of $\tilde{T}_j$ (j=1,2), the Kaplan-Meier estimate assigns mass $p^j(t_k^j) =: p_k^j$ to all K_j observed ordered event times t_k^j, $k = 1, ..., K_j$. The estimate is at least available up to the largest observation of T_j, leaving some mass $p^j_{K_j+1}$ un-specified if the largest observation is censored. Both estimates are discrete up to $\tilde{m}_{1,2} := \min(\max T_1, \max T_2)$.

For $t_* \leq t^* \leq \tilde{m}_{1,2}$, we can estimate $P(t_* \leq \tilde{T}_{min} \leq t^*)$, while assuming $\tilde{T}_1$ and $\tilde{T}_2$ independent, by

$$p_0(t_*, t^*) := \sum_{t_* \leq t_k^1 \leq t^*} \big(p_k^1 (1 - \sum_{l < t_k^1} p^2(l)) \big) + \sum_{t_* \leq t_k^2 \leq t^*} \big(p_k^2 (1 - \sum_{l \leq t_k^2} p^1(l)) \big).$$

With $\delta_0^* := \tilde{m}_{1,2} - t^*$, we obtain an estimate of $\mathfrak{D}_0$, which is discrete up to δ_0^* and assigns mass

$$q_{d_0} := 1/p_0(t_*, t^*) \cdot \sum_{t_* \leq t_k^1 \leq t^*} \sum_{t_* \leq t_l^2 \leq t^*} p_k^1 \cdot p_l^2 \cdot \mathbb{1}_{\{|t_k^1 - t_l^2| = d_0\}}(t_k^1, t_l^2)$$

to any value $d_0 \in \{d_1^0, ..., d_M^0\}$, where all possible absolute difference $0 \leq |t_k^1 - t_l^2| \leq \delta_0^*$ with $t_* \leq \min(t_k^1, t_l^2) \leq t^*$ are included. Note that this set covers all values in $\{d_1, ..., d_L\}$.

4.2 Relative measures

To compare closeness across word pairs we consider in the following some potential candidates for relative measures $r(\mathfrak{D}, \mathfrak{D}_0)$. Such measures can simply be constructed by calculating a (suitable) absolute measure (cf. Table 1) for both $\mathfrak{D}$ and $\mathfrak{D}_0$. Then, the two are set into perspective by calculating $r(\mathfrak{D}, \mathfrak{D}_0) = f(m(\mathfrak{D}), m(\mathfrak{D}_0))$ with an appropriate function f, such as the difference or ratio. Examples are $r_0 := m_0(\mathfrak{D}) - m_0(\mathfrak{D}_0)$ or $r_x := m_x(\mathfrak{D}) - m_x(\mathfrak{D}_0)$ with $x = 1, 2, 3$.

In this paper, we will focus on measures of this type. Note, however, that not every suitable measure can be expressed this way. For example, one might be interested in

$$r_\alpha := P\,(\tilde{\Delta} \leq F_{\mathfrak{D}_0}^{-1}(\alpha)),$$

with $\alpha = 0.25$, say. In our data, this approach might improve comparisons across pairs, where "small" means different things, but it is also more complicated since it requires to specify t^* very carefully. See also section 6 for another alternative approach.

To compute relative measures which depend on x, such as r_x, the estimate of $\mathfrak{D}_0$ has to be discrete up to x. This may imply that we have to choose t^* smaller than $m_{1,2}^*$, that is, smaller than necessary to estimate $\mathfrak{D}$ alone, to ensure that $x \leq \delta_0^*$ (cf. our example in section 7). Note, however, that δ_0^* might be larger than δ^*. In this case, the problem vanishes, since $x \leq \delta_0^*$ holds, provided that $x \leq \delta^*$.

4.3 Relative measures conditioned on $\tilde{T}_{min}$

If $\mathfrak{D}_t$ depends heavily on t, one may argue that we get a sharper contrast by comparing $\mathfrak{D}_t$ to $\mathfrak{D}_{0t}$, the corresponding conditional distribution under independence of $\tilde{T}_1$ and $\tilde{T}_2$. This suggests to consider relative measures of the type

$$E\left(r(\mathfrak{D}_{\tilde{T}_{min}}, \mathfrak{D}_{0\tilde{T}_{min}}) \,|\, t_* \leq \tilde{T}_{min} \leq t^* \right).$$

Estimation is straightforward:

1. Since the values p_k (cf. section 3.1) provide us with a discrete estimate of $\tilde{T}_{min}$ given $t_* \leq \tilde{T}_{min} \leq t^*$, we can estimate this measure by

   $$\sum_{t_* \leq t_k \leq t^*} r(\hat{\mathfrak{D}}_{t_k}, \hat{\mathfrak{D}}_{0t_k}) \cdot p_k$$

   with suitable estimates $\hat{\mathfrak{D}}_{t_k}$ and $\hat{\mathfrak{D}}_{0t_k}$ (provided that the relative measure only relies on estimates up to $\min(\delta^*, \delta_0^*)$).

2. $\mathfrak{D}_{t_k}$ is estimated as described in section 3.2 based on a Cox proportional hazards model, yielding a discrete estimate $\hat{\mathfrak{D}}_{t_k}$ (at least) up to δ^*, which assigns probability mass $\tilde{q}_d(k)$ to each observed $\tilde{\Delta}$-value d in $d_1, ..., d_L$.

3. For $t_* \leq t_k \leq t^*$ and $0 < d_0 \leq \delta_0^*$, we can obtain an estimate of $\mathfrak{D}_0$, putting mass

   $$\frac{p^1(t_k)\, p^2(t_k + d_0) \;+\; p^1(t_k)\, p^1(t_k + d_0)}{p^1(t_k)\left(1 - \sum_{l<t_k} p^2(l)\right) + p^2(t_k)\left(1 - \sum_{l \leq t_k} p^1(l)\right)} \tag{1}$$

   on each $d_0 \in \{d_1^0, ..., d_M^0\}$ (cf. section 4.1). For $d_0 = 0$, this modifies to

   $$\frac{p^1(t_k)\, p^2(t_k)}{p^1(t_k)\left(1 - \sum_{l<t_k} p^2(l)\right) + p^2(t_k)\left(1 - \sum_{l \leq t_k} p^1(l)\right)}. \tag{2}$$

See also section 5 below.

5 Taking a covariate S into account

It may be the case that we know one or more additional factors which can explain (some) closeness since the factor(s) are correlated with both $\tilde{T}_j$ (j=1,2). In the example, the variation of single acquisition times might be smaller within groups of children sharing a certain factor level $S = s$ than across groups. Such an influential factor S could be, for example, the children's general "speed" in building up their vocabulary. In such a case, we would be interested in quantifying (only) the *excess* closeness, that is, the closeness which cannot be explained by the joint correlation with S. We can approach this by choosing $\mathfrak{D}_0$ as the distribution of $\tilde{\Delta}$ under the assumption of independence of $\tilde{T}-1$ and $\tilde{T}_2$ given S, such that the amount of closeness under chance condition increases.

In the following, censoring assumptions (C1)-(C4) have to be modified appropriately, taking S into account.

The basic idea is as follows: We consider for suitable $t_* \leq t \leq t^*$ the distribution

$$\mathfrak{D}_{0t}(s) := \mathfrak{D}_{\tilde{T}_1 \perp \tilde{T}_2 | S=s}(\tilde{\Delta} \,|\, S = s,\, \tilde{T}_{min} = t),$$

and compare this to $\mathfrak{D}_t(s) := \mathfrak{D}(\tilde{\Delta} \,|\, S = s,\, \tilde{T}_{min} = t)$ forall t. Then we can define a relative measure of closeness as

$$E\left(r\big(\mathfrak{D}_{\tilde{T}_{min}}(S), \mathfrak{D}_{0\tilde{T}_{min}}(S)\big) \,|\, t_* \leq \tilde{T}_{min} \leq t^* \right).$$

Estimation can be done along the following steps:

1. With p_k from section 3.1, we can estimate this measure by

$$\sum_{t_* \leq t_k \leq t^*} \big(\, \text{ave}_{i, \tilde{T}_{min}=t_k} \, r(\hat{\mathfrak{D}}_{t_k}(s_i), \hat{\mathfrak{D}}_{0t_k}(s_i)) \,\big) \cdot p_k$$

   with suitable estimates $\hat{\mathfrak{D}}_{t_k}$ and $\hat{\mathfrak{D}}_{0t_k}$ (provided that the relative measure only relies on estimates up to $\min(\delta^*, \delta_0^*)$).

2. Extending the considerations of section 3.2 for t alone, $\mathfrak{D}_{t_k}(s)$ can be estimated based on a

Cox proportional hazards model, that is

$$\lambda(d \,|\, t, s) = \lambda_{\text{baseline}}(d) \cdot g_\beta(t, s),$$

where a simple choice of $g_\beta(t,s)$ may be $\exp(\beta_t \cdot t + \beta_s \cdot s)$. This yields an estimate of $\mathfrak{D}_{t_k}(s_i)$ which is discrete (at least) up to δ^*, assigning a probability to each value d in $\{d_1, ..., d_L\}$.

3. Estimation of $\mathfrak{D}_{0t_k}(s)$ requires to have estimates of the conditional distribution of both $\tilde{T}_j$ (j = 1,2) given $S = s$. Such estimates can again be obtained from a Cox proportional hazards models (for each $\tilde{T}_j$), given by:

$$\lambda(t \,|\, s) = \lambda_{\text{baseline}}(t) \cdot g_\beta(s),$$

where the baseline hazard is estimated according to Kalbfleisch & Prentice (2002: 115). For example, the function $g_\beta(t)$ could be defined as $\exp(\beta \cdot s)$. Modeling yields discrete estimates, assigning probability masses to all observed event times t_k^j. Based on these probability masses, $\mathfrak{D}_{0t_k}(s_i)$ can be estimated correspondingly as outlined in eq.(1) and eq.(2) of section 4.3, at least up to δ_0^*. Note that this property may be lost, if one uses stratification with respect to S or nearest-neighbour-smoothing.

If the covariate we would like to take into account is time-dependent, that is, if the value $S(U) = s$ of a factor is measured at the (random) time point U, we have to be careful in defining the reference distribution $\mathfrak{D}_0$, since we have the choice between ($t_* \leq t \leq t^*$):

$$\mathfrak{D}^1_{0t}(u, s) := \mathfrak{D}_{\tilde{T}_1 \perp \tilde{T}_2 | S(u)=s}(\tilde{\Delta} \,|\, \tilde{T}_{min} = t, S(u) = s),$$

or

$$\mathfrak{D}^2_{0t}(s) := \mathfrak{D}_{\tilde{T}_1 \perp \tilde{T}_2 | S(U)=s}(\tilde{\Delta} \,|\, \tilde{T}_{min} = t, S(U) = s).$$

Here, we would like to consider the quantity

$$E\left(r\big(\mathfrak{D}_{\tilde{T}_{min}}(S(U)), \mathfrak{D}^1_{0\tilde{T}_{min}}(U, S(U))\big) \,|\, t_* \leq \tilde{T}_{min} \leq t^* \right)$$

or

$$E\left(r\big(\mathfrak{D}_{\tilde{T}_{min}}(S(U)), \mathfrak{D}^2_{0\tilde{T}_{min}}(S(U))\big) \,|\, t_* \leq \tilde{T}_{min} \leq t^* \right),$$

which we can be estimated as above by

$$\sum_{t_* \leq t_k \leq t^*} \big(\, \text{ave}_{i,\tilde{T}_{min}=t_k}\; r(\hat{\mathfrak{D}}_{t_k}(S(u_i)), \hat{\mathfrak{D}}^*_{0t_k}(u_i, S(u_i)))\,\big) \cdot p_k$$

or

$$\sum_{t_* \leq t_k \leq t^*} \big(\, \text{ave}_{i,\tilde{T}_{min}=t_k}\; r(\hat{\mathfrak{D}}_{t_k}(S(u_i)), \hat{\mathfrak{D}}^*_{0t_k}(S(u_i)))\,\big) \cdot p_k,$$

respectively, with suitable estimates $\hat{\mathfrak{D}}_{t_k}(s)$ and $\hat{\mathfrak{D}}^1_{0t_k}(u, s)$ or $\hat{\mathfrak{D}}^2_{0t_k}(s)$.

Estimation of $\mathfrak{D}_{t_k}(s)$ can be approached as above, replacing S by $S(U)$ or $S(u)$, respectively. Estimation of $\hat{\mathfrak{D}}^2_{0t_k}(s)$ requires to model $\mathfrak{D}(\tilde{T}_j \,|\, S(U) = s)$ and estimation of $\hat{\mathfrak{D}}^1_{0t_k}(u, s)$ requires to model $\mathfrak{D}(\tilde{T}_j \,|\, S(u) = s)$, which can in principle be approached as described above in the case of time-independent covariates.

The only additional difficulty may arise from the fact that, due to censoring at C^U, U – and therefore $S(U)$ or $S(u)$ – might be unknown for some subjects. Omission of these subjects in modeling $\mathfrak{D}_{t_k}(s)$ requires to assume that (1) $\tilde{\Delta}$ and $C^U \geq u$ are conditionally independent given $\tilde{T}_{min} = t, S(u) = s$ or (2) $\tilde{\Delta}$ and $C^U \geq U$ are conditionally independent given $\tilde{T}_{min} = t, S(U) = s$ ($t_* \leq t \leq t^*$), which is for example satisfied if $U = \tilde{T}_{min}$. In modeling $\mathfrak{D}(\tilde{T}_j \,|\, S(U) = s)$ (or $\mathfrak{D}(\tilde{T}_j \,|\, S(u) = s)$), omission of these subjects requires additionally to assume that $\tilde{T}_j$ is conditionally independent of $C^U \geq U$ (or $C^U \geq u$) given $S(U) = s$ (or $S(u) = s$).

If one does not want to consider the relative measure conditioned on $\tilde{T}_{min}$, $\mathfrak{D}_{\tilde{T}_{min}}$ and $\mathfrak{D}_{0\tilde{T}_{min}}$ can be replaced in the definitions above by $\mathfrak{D}$ and $\mathfrak{D}_0$, respectively. Then, the estimation simplifies since we do not have to take the weighted average in step 3.

6 An alternative approach to a relative measure

Instead of defining a relative measure based on estimates for both $\mathfrak{D}$ and $\mathfrak{D}_0$, one could employ the following "short-cut" hazard approach which is based on an (assumed) constant hazard ratio (up to a value x) between the two distributions.

Consider the hazard function $\lambda(d\,|\,t)$ of $\mathfrak{D}_t$ and the corresponding hazard function $\lambda_0(d\,|\,t)$ of $\mathfrak{D}_{0t}$, ignoring a potential covariate S for the moment. We assume that the ratio is constant for all $t_* \leq t \leq t^*$ up to a pre-specified value x:

$$\frac{\lambda(d\,|\,t)}{\lambda_0(d\,|\,t)} = \theta \text{ for } 0 \leq d \leq x.$$

The distribution of

$$\tilde{\Delta}* = \begin{cases} \tilde{\Delta} & \text{if } \tilde{\Delta} \leq x \\ x+1 & \text{otherwise} \end{cases} \text{, given } \tilde{T}_{min} = t,\ t_* \leq t \leq t^*$$

is now completely specified by $\lambda_0(d\,|\,t)$ and θ. Note that the value $x+1$ is completely arbitrary.

With $d_i = \min(\tilde{\Delta}_i, C_i^{\Delta})$, define an event indicator for each subject i with $t_* \leq \tilde{T}_{min} = t_i \leq t^*$ as follows:

$$\delta_i = \begin{cases} 1 & \text{if } d_i \leq \min\,(C_i^{\Delta}, x) : \tilde{\Delta} \text{ observed as } d_i \leq x \\ 0 & \text{if } d_i > \min\,(C_i^{\Delta}, x). \end{cases}$$

With $d_i^* = \min(d_i, x)$, let $\{d_{[l]}\}$ $(l = 1, ..., M_i)$ be the ordered subset of $\{d_l\}$ $(l = 1, ..., M)$, which are smaller or equal than d_i^*. If $x \leq \delta_0^*$, we can translate our discrete estimate for $\mathfrak{D}_{0t_i}$ (see section 4.3, in particular eq.(1) and eq.(2)) into estimates of $\lambda_0(d_{[l]}\,|\,t_i)$ $(l = 1, ..., M_i)$. Note that $\hat{\lambda}_0(d_i\,|\,t_i) \neq 0$ for all i with $\delta_i = 1$.

The contribution for subject i to the likelihood of $\tilde{\Delta}^*$ can then be approximated by:

$$c\,(i,\theta) := \begin{cases} P\,(\tilde{\Delta}^* = d_i^* \,|\, \tilde{T}_{min} = t_i,\, \theta) = \theta\lambda_0(d_i \,|\, t_i) \cdot \prod_{l=1}^{i-1} (1 - \theta\lambda_0(d_{[l]} \,|\, t_i)) & \text{if } \delta_i = 1 \\ P\,(\tilde{\Delta}^* > d_i^* \,|\, \tilde{T}_{min} = t_i,\, \theta) = \prod_{l=1}^{i} (1 - \theta\lambda_0(d_{[l]} \,|\, t_i)) & \text{if } \delta_i = 0. \end{cases} \quad (3)$$

Now we suggest to estimate θ by maximizing the approximate likelihood:

$$\prod_{i,\, t_* \leq \tilde{T}_{min,i} \leq t^*,\, \tilde{T}_{min} \leq C} c\,(i,\theta) \quad (4)$$

in θ by plugging the estimates of $\lambda_0(d \,|\, t)$ in eq.(3). The restriction to subjects where the minimum of the two ecvent times is observed is again justified by assumption (C4).

Adjustment with respect to covariates S can be accomplished by expanding the assumption of a constant ratio (for all $t_* \leq t \leq t^*$) to:

$$\frac{\lambda(d \,|\, t, s)}{\lambda_0(d \,|\, t, s)} = \theta \text{ for } 0 \leq d \leq x \text{ and all } s.$$

Replacing $\lambda_0(d \,|\, t)$ by $\lambda_0(d \,|\, t, s)$ in defining $c\,(i,\theta)$, eq.(4) can be maximized in θ based on estimates for $\lambda_0(d \,|\, t, s)$, which can be derived as outlined in section 5, item 3.

Note that this approach has the additional advantage of only requiring estimation of $\mathfrak{D}_{0t}$ or $\mathfrak{D}_{0t}(s)$, but not of $\mathfrak{D}_t$ or $\mathfrak{D}_t(s)$. Note also that by applying the usual inference techniques for ML estimates, we can obtain standard errors for $\hat{\theta}$ and confidence intervals. However, these do not take into account the uncertainty of the results due to estimation of $\mathfrak{D}_0$.

7 Example

7.1 Analysis overview

From the Danish Longitudinal CDI study (Andersen et al. 2006: chapter 3; Wehberg et al. 2007), acquisition times of 410 different words can be defined for the 182 children who participated in the study. Children are observed monthly at the beginning of each month from 8 to at most 30 months of age, giving rise to events T_{ij} (= time point of first reported occurrence of a word j for child i = acquisition time) at time points: $t = 9, ..., z_i$, where $z_i \leq 30$ denotes the last observed month of child i. Information on words not reported until z_i is right-censored – the word could only have been acquired later, if at all – and information on words reported at time point 8 is left-censored (cf. Table 2). Some families failed to contribute CDI forms in some months (before resuming study participation), mostly due to vacation months or the like, resulting in some interval-censored observations. Since we typically have for each word t_j only a few children who are affected by this (real) interval-censoring (cf. Figure 1 and Table 3), we perform for each pair of words a complete case analysis, that is, excluding the few children where one or both acquisition times are affected by interval-censoring.

Table 2. Left-censoring: number of children N where a specific word already appears as said in month 8

Word	Translation	N
grrrr	(lion sound)	4
hej	hi	3
mm mm (lækkert)	(tastes-well sound)	3
mad	food/mealtime	2
hej hej (farvel)	bye-bye	1
ja	yes	1
mor	mother	1
nej	no	1

Figure 1. Number of children N per word where the acquisition times lie in an interval

Table 3. Words where the acquisition times lie in an interval for more than 14 children

Word	Translation	N of children	N of events after omitting
aftensmad	evening meal / supper	15	140
blå	blue	15	136
kjole	dress	15	134
knæ	knee	15	154
spaghetti	spaghetti / pasta	15	126
tænder	teeth	15	150
knap	button	18	137

For all analyses, we set the lower border $t_* = 9$, which allows us to ignore that observations with $T_{ij} = 8$ are actually left-censored when estimating. The upper border $t^* = 26$ allows us to estimate $\mathfrak{D}_0$ up to 3, since for all words (but one) the largest observation is either censored at 29 or 30 months or not censored at all. We exclude the one word (*far (father)*), where the largest observation is censored at 28 months. Further, we exclude 13 words with median acquisition times above 30 (cf. Table 4), leaving only words with more than 70 events (78,210 pairs). Since we expect closeness only for pairs of words which are on average learnt rather close, we restrict analyses to 61,602 pairs where the distance between the (estimated) medians is smaller than 6 months (excluding 16,608 pairs). In all analyses, we use the simple linear Cox models, which are

Table 4. Words with less than 71 events

Word	Translation	N of events
intet	nothing	35
person	person	39
morbror	mother's brother	44
farbror	father's brother	48
tante	aunt	52
hendes	her	57
lærer	teacher	58
køn	sex/gender	61
småpenge	coins	61
(et) andet	something else	62
hvordan	how?	65
park	park	67
sweater	sweater	70

mentioned as examples above. All analyses are done in Stata™.

Three different analyses are performed. First, we investigate which pairs are close, based on the absolute measures m_0 and m_2. Table 5 presents the results for the 30 closest pairs. The ranking for m_0 is done as follows: all 61,602 pairs are sorted by m_0 (if necessary, followed by sorting by m_2 and alphabetically by word 1 and 2), ranking for m_2 is done correspondingly. Second, we look at the results based on relative measures: Table 6 presents results for the 30 closest pairs of r_0 and r_2 based on the difference ($diff_0$ and $diff_2$), and Table 7 presents corresponding results based on the ratio ($ratio_0$ and $ratio_2$). The alternative hazard approach (HA) is illustrated in Table 8. Here, ranking is not done by the estimated parameter θ, but the lower bound of the confidence interval for θ instead.

Third, we take two different covariates into account: gender (88 boys = 48%; 94 girls = 52%) and (continuous) level of communicative ability. Considering gender is motivated by the fact that there exist significant differences in acquisition speed and in the content of first words between boys and girls (cf. Wehberg et al. 2007). Communicative ability is measured as $S(\tilde{T}_{min})$ = Infant lexicon size at the beginning of the month where the first of the two words is learnt. Since the aim is to compare each child with children having a similar vocabulary size at $\tilde{T}_{min}$, we condition on $S(\tilde{T}_{min})$ and not on $S(t_{min})$. Here, it seems possible to assume that the censoring time C is independent of the single event times $\tilde{T}_1$ and $\tilde{T}_2$, given $S(\tilde{T}_{min})$: if C depends on the individual development at all, it is reasonable to assume that it is related to the overall size of the lexicon, that is, parents may be more inclined to stop filling in the questionnaires when the child approaches the upper limit of the CDI vocabulary checklist. However, it is unlikely that C depends additionally on the outcome of single/specific items. Table 9 presents an illustration of the results based on the hazard approach.

7.2 Results

To ease readability, only the English translations of the Danish words will be used in the following. In the tables, both Danish words and their English translation are presented. Note that the translated words do not necessarily match items/words in the original American CDI vocabulary checklist.

Word pairs, which appear close under the absolute measure m_0 (Table 5), share some interesting features: besides words which occur twice (in different categories) in the CDI vocabulary list (e.g., *fish* occurs both in the category *Animal names* and *Food and drink*), many pairs consist of body parts such as *arm*, *leg*, or *nose*. Other pairs are clear opposites, e.g., *no-yes* or *boy-girl*, or tend to have a strong association (*red-blue*). The results for m_2 are similar: at least 80% in both lists have ranks smaller than 300 under the other measure. One interesting example, though, is *grandfather-grandmother*, which has the second rank under m_0 but only rank 318 under m_2: this is a clear indication that if the two words are acquired close together, they are typically acquired in the same month. Another difference is that pairs of body parts are even more pronounced under m_2, whereas opposing pairs seem to lose slightly in rank.

To gain a first impression on the relation between more than two words at one time, we performed a hierarchical cluster analysis for both similarity measures m_0 and m_2. Based on m_0, for example, there are 20 clusters of size ≥ 3 with an average pairwise value of $m_0 \geq 0.25$: the largest cluster consists of 6 words (*hand, leg, bed, arm, head, finger*) and the next largest of 5 words (*hair, eye, nose, mouth, ear*). All other clusters contain 4 (e.g., *tiger, giraffe, elephant, lion*) or 3 words (e.g., *to fall, dirty, to run*). Not surprisingly, results for m_2 are similar: for example, the 2 clusters of size ≥ 3 with an average pairwise value of $m_2 \geq 0.8$ resemble closely the two largest cluster for m_0: *hand, leg, bed, arm, head, finger* plus *fork* and *hair, eye, nose, mouth, ear* minus *hair*.

Relative measures based on the difference are presented in Table 6. The resulting pairs are very similar to those based on the absolute measures, especially for diff_0: the comparison to $\mathfrak{D}_0$ based on the difference does not result in the intended re-assessment of closeness amongst the pairs, as can be seen by the corresponding, relatively high ranking based on m_0. Taking into account m_0^0, the absolute measure under $\mathfrak{D}_0$, one sees clearly why: for most pairs, m_0^0 is very small, and consequently, the difference is not much different from the absolute measure. The situation for diff_2 seems slightly better, since m_2^0 itself is larger and the differences between the pairs become more pronounced. But considering the ratio instead of the difference seems a more natural choice and, not surprisingly, leads to a more drastic re-assessment compared to the absolute measures (Table 7).

Regarding the word pairs which are on top of the lists under ratio_0 and ratio_2, the content of the words (such as being a body part) does not seem to be the most obvious common factor any longer. Again, we performed a hierarchical cluster analysis for both similarity measures ratio_0 and ratio_2. Based on ratio_2, for example, there are 4 clusters of size ≥ 3 with an average pairwise value of $\mathit{ratio}_2 \geq 3$: the largest cluster consists of 10 words (*to watch, (small) box, playpen, rocking chair, nice, pleasant/nice, person/human being, shop/business, sledge, to stop*) and the two next largest of 4 words: *to sit, to show, to hurry, behind* and *deer/stag, pony, turkey, puppy.* The smallest cluster here is of size 3: *to dry, to pull, to take.* Note, however, that Table 7 points to a methodological problem: when the absolute measure based on $\mathfrak{D}0$ becomes close to zero, the resulting ratios are very unstable and highly sensitive to small random disturbances of the absolute measure under $\mathfrak{D}0$.

An illustration of the alternative relative measure, the hazard approach *HA*, is presented in Table 8[1]. Note that for 1,358 of the 61,602 pairs, the ML algorithm did not converge. To base ranking on the lower bound of the confidence interval instead of the estimated coefficient itself ensures that we do not over-interpret an insecure, high θ. As we can see, the ranking after θ itself differs from the ranking after the confidence interval, but not dramatically. Cluster analysis (where a missing value for the lower bound of the confidence interval for θ is interpreted as the largest possible distance for the pair) supports the similarity of the results to those based on $ratio_2$: for example, there are 6 clusters of size ≥ 3 with an average pairwise value of the lower bound ≥ 5: the largest cluster consists of 8 words (*church, zoo, playpen, highchair, shop/business, toboggan/sledge, deer/stag, to stop*), and the next largest of 4 words: *bad, sledge, to end, good.* The items *person/human being, rocking chair, pony, to sit*, for example, are contained in the 4 other clusters of size 3.

As mentioned above, conditioning on $\tilde{T}_{min}$ influences the results but slightly when using a relative measure based on the difference. Calculating a relative measure based on the ratio instead, encounters the problem that for some t, the denominator might be (very) small or actually zero, inflating the (averaged) ratio disproportionately, especially when conditioning on $\tilde{T}_{min} = t$. Also the hazard approach has some methodological issues which become clear when taking covariates into account (see Table 9 for illustration purposes). Take for example *grandfather-grandmother.* Conditioning (arbitrarily) on $\tilde{T}_{min} = 18$, the estimated hazard is 0.023 at zero and increases to 0.055 at 3 under $\mathfrak{D}_0$, while it is 0.445 at zero under $\mathfrak{D}$, decreasing to 0.189 at 3. Assuming a constant hazard ratio up to 3 yields an estimate of 8.5, which results in an estimated hazard

[1]Unfortunately, I discovered a mistake in the implementation of this approach too late to re-calculate the results for all 61,602 pairs. For Table 8, analysis of those 1743 pairs, which were either among the first 1000 pairs sorted by the original θ, the first 1000 pairs sorted by the original lower bound of the confidence interval for θ, or the first 1000 pairs sorted by $ratio_2$, was re-done with similar results for most pairs. Results of the cluster analysis reported here are based on the original analysis.

0.194 at zero (under $\mathfrak{D}$) and 0.470 at 3. It becomes clear that the choice of x is crucial to obtain interpretable results, especially when the estimated hazards under $\mathfrak{D}_0$ are close to zero. Taking gender into account, the estimated θ drops to 4.8, which seems a relevant decrease implying a large gender effect. Looking more closely at the estimated hazards, however, puts this result in perspective. Even though the estimates based on $\mathfrak{D}_0$ are not directly comparable, since now they are based on Cox modeling instead of Kaplan-Meier estimates, the actual numbers are quite similar: for girls (boys), the estimated hazard is 0.035 (0.033) at zero and increases to 0.086 (0.068) at 3 under $\mathfrak{D}_0$, while it is 0.462 (0.427) at zero under $\mathfrak{D}$, decreasing to 0.200 (0.181) at 3. Plugging-in the estimated $\theta = 4.8$, results in an estimated hazard 0.168 (0.159) at zero (under $\mathfrak{D}$) and 0.415 (0.396) at 3. This illustrates nicely how (very) small changes in the estimates of $\mathfrak{D}_0$ can have a large effect on the estimated θ and suggests that the hazard approach as well may benefit from some refinement.

Table 5. Comparison based on the absolute measures m_0 and m_2. In addition, ranking R after the other measure is presented.

	m_0				m_2			
Rank	Pair		m_0	R m_2	Pair		m_2	R m_0
1	arm	ben (leg)	0.592	1	arm	ben (leg)	0.945	1
2	bedstefar (grandfather)	bedste(mor) (grandmother)	0.574	318	finger	hånd (hand)	0.940	15
3	øje (eye)	øre (ear)	0.557	5	arm	hånd (hand)	0.932	21
4	fisk (fish)	fisk	0.509	9	tænder (teeth)	finger	0.912	132
5	vand (water)	vand	0.488	10	øje (eye)	øre (ear)	0.911	3
6	gynge (swing)	gynge (to swing)	0.460	27	hoved (head)	hånd (hand)	0.908	101
7	næse (nose)	øje (eye)	0.450	18	ben (leg)	hånd (hand)	0.896	26
8	mund (mouth)	øre (ear)	0.435	16	arm	finger	0.890	38
9	kylling (chicken)	kylling	0.426	22	fisk (fish)	fisk	0.885	4
10	næse (nose)	øre (ear)	0.425	23	vand (water)	vand	0.884	5
11	mund (mouth)	næse (nose)	0.412	21	ben (leg)	finger	0.881	52
12	rød (red)	blå (blue)	0.411	71	arm	hoved (head)	0.871	126
13	mund (mouth)	hår (hair)	0.410	45	strømper (stockings/socks)	bukser (trousers)	0.869	78
14	bluse (blouse)	bukser (trousers)	0.410	34	finger	hoved (head)	0.862	90
15	finger	hånd (hand)	0.404	2	finger	seng (bed)	0.862	96
16	mund (mouth)	øje (eye)	0.398	25	mund (mouth)	øre (ear)	0.860	8
17	hej (hi)	hej hej (farvel) (bye-bye)	0.394	401	finger	gaffel (fork)	0.858	811
18	pude (pillow)	dyne (blanket)	0.386	105	næse (nose)	øje (eye)	0.858	7
19	nej (no)	ja (yes)	0.385	297	ben (leg)	hoved (head)	0.857	310
20	sætte (to sit)	vise (to show)	0.385	394	håndklæde (towel)	køleskab (refrigerator)	0.850	35
21	arm	hånd (hand)	0.368	3	mund (mouth)	næse (nose)	0.847	11
22	dreng (boy)	pige (girl)	0.366	147	kylling (chicken)	kylling	0.847	9
23	nattøj (pyjamas)	køkken (kitchen)	0.354	48	næse (nose)	øre (ear)	0.845	10
24	øje (eye)	hår (hair)	0.352	115	arm	dyne (blanket)	0.844	1362
25	hund (dog)	kat (cat)	0.349	3076	mund (mouth)	øje (eye)	0.844	16
26	ben (leg)	hånd (hand)	0.348	7	bukser (trousers)	arm	0.841	129
27	øre (ear)	hår (hair)	0.345	57	gynge (swing)	gynge (to swing)	0.841	6
28	løbe (to run)	falde (to fall)	0.343	28	løbe (to run)	falde (to fall)	0.839	28
29	barnevogn (baby carriage)	klapvogn (baby buggy)	0.341	373	tænder (teeth)	briller (glasses)	0.836	5025
30	lukke (to shut)	åbne (to open)	0.341	321	spise (to eat)	finger	0.835	169

Table 6. Comparison based on the relative measures *diff*$_0$ and *diff*$_2$. Corresponding absolute measures m_x under $\mathfrak{D}_0$ are included as m_x^0. In addition, ranking R after m_0 and m_2, respectively, is presented.

	diff$_0$					*diff*$_2$				
Rank	Pair		*diff*$_0$	m_0^0	R m_0	Pair		*diff*$_2$	m_2^0	R m_2
1	bedstefar (grandfather)	bedste(mor) (grandmother)	0.543	0.031	2	bedstefar (grandfather)	bedste(mor) (grandmother)	0.623	0.142	318
2	arm	ben (leg)	0.494	0.097	1	forretning (shop/business)	kravlegård (playpen)	0.559	0.157	1263
3	øje (eye)	øre (ear)	0.471	0.086	3	finger	hånd (hand)	0.553	0.388	2
4	fisk (fish)	fisk	0.426	0.083	4	tage (to take)	tørre af (to dry)	0.552	0.220	239
5	vand (water)	vand	0.403	0.085	5	håndklæde (towel)	køleskab (refrigerator)	0.551	0.300	20
6	gynge (swing)	gynge (to swing)	0.393	0.067	6	tænder (teeth)	finger	0.537	0.376	4
7	næse (nose)	øje (eye)	0.365	0.085	7	soveværelse (bedroom)	badeværelse (bathroom)	0.537	0.262	101
8	kylling (chicken)	kylling	0.356	0.070	9	sætte (to sit)	vise (to show)	0.533	0.225	394
9	mund (mouth)	øre (ear)	0.351	0.084	8	ren (clean)	tør (dry)	0.527	0.254	181
10	rød (red)	blå (blue)	0.351	0.060	12	sætte (to sit)	ren (clean)	0.526	0.223	496
11	sætte (to sit)	vise (to show)	0.341	0.043	20	arm	ben (leg)	0.525	0.421	1
12	næse (nose)	øre (ear)	0.340	0.085	10	arm	hånd (hand)	0.524	0.409	3
13	bluse (blouse)	bukser (trousers)	0.336	0.074	14	kasse (box)	kost (broom)	0.523	0.305	39
14	mund (mouth)	hår (hair)	0.328	0.082	13	gynge (swing)	gynge (to swing)	0.519	0.321	27
15	mund (mouth)	næse (nose)	0.326	0.086	11	æske (small box)	kravlegård (playpen)	0.516	0.165	2728
16	finger	hånd (hand)	0.324	0.080	15	hoved (head)	hånd (hand)	0.514	0.394	6
17	mund (mouth)	øje (eye)	0.313	0.086	16	skynde sig (to hurry)	tørre af (to dry)	0.513	0.217	887
18	pude (pillow)	dyne (blanket)	0.308	0.078	18	strømper (stockings/socks)	bukser (trousers)	0.511	0.358	13
19	hej (hi)	hej hej (farvel) (bye-bye)	0.299	0.095	17	håndklæde (towel)	soveværelse (bedroom)	0.511	0.290	87
20	dreng (boy)	pige (girl)	0.292	0.074	22	forretning (shop/business)	standse (to stop)	0.510	0.169	2892
21	pæn (nice)	rar (pleasant/nice)	0.287	0.028	41	øje (eye)	øre (ear)	0.510	0.401	5
22	nattøj (pyjamas)	køkken (kitchen)	0.286	0.067	23	sætte (to sit)	tørre af (to dry)	0.507	0.211	1216
23	gyngestol (rocking chair)	kravlegård (playpen)	0.286	0.026	47	vise (to show)	bagved (behind)	0.505	0.217	1109
24	forretning (shop/business)	holde øje med (to watch)	0.285	0.029	43	køkken (kitchen)	køleskab (refrigerator)	0.503	0.315	49
25	nej (no)	ja (yes)	0.283	0.102	19	skynde sig (to hurry)	trække (to pull)	0.503	0.229	867
26	hund (dog)	kat (cat)	0.282	0.067	25	vand (water)	vand	0.502	0.382	10
27	arm	hånd (hand)	0.281	0.088	21	skuffe (drawer)	soveværelse (bedroom)	0.502	0.281	169
28	rydde op (to clean up)	smide (to throw)	0.276	0.053	34	kylling (chicken)	kylling	0.501	0.345	22
29	lukke (to shut)	åbne (to open)	0.275	0.065	30	rød (red)	blå (blue)	0.500	0.307	71
30	håndklæde (towel)	køleskab (refrigerator)	0.275	0.051	35	sige (to say)	skuffe (drawer)	0.500	0.264	336

Table 7. Comparison based on the relative measures *ratio*$_0$ and *ratio*$_2$. Corresponding absolute measures m_x under $\mathfrak{D}_0$ are included as m_x^0. Additional ranking R is presented.

	ratio$_0$						*ratio*$_2$					
Rank	Pair		*ratio*$_0$	m_0^0	R m_0	R *diff*$_0$	Pair		*ratio*$_2$	m_2^0	R m_2	R *diff*$_2$
1	bedstefar (grandfather)	bedste(mor) (grandmother)	18.580	0.031	2	1	bedstefar (grandfather)	bedste(mor) (grandmother)	5.396	0.142	318	1
2	holde øje med (to watch)	standse (to stop)	11.914	0.022	318	63	slæde (sledge)	menneske (person/human being)	4.593	0.135	7594	41
3	gyngestol (rocking chair)	kravlegård (playpen)	11.907	0.026	47	23	forretning (shop/business)	kravlegård (playpen)	4.562	0.157	1263	2
4	menneske (person/human being)	standse (to stop)	11.374	0.023	283	61	menneske (person/human being)	holde øje med (to watch)	4.156	0.153	6274	47
5	pæn (nice)	rar (pleasant/nice)	11.274	0.028	41	21	menneske (person/human being)	kravlegård (playpen)	4.152	0.151	6979	49
6	hvorfor (why?)	hvornår (when?)	11.202	0.016	5231	946	slæde (sledge)	forretning (shop/business)	4.147	0.139	13137	197
7	slem (bad)	hvornår (when?)	11.164	0.022	529	93	æske (small box)	kravlegård (playpen)	4.122	0.165	2728	15
8	forretning (shop/business)	holde øje med (to watch)	10.836	0.029	43	24	forretning (shop/business)	standse (to stop)	4.012	0.169	2892	20
9	hvalp (puppy)	pony	10.767	0.028	73	33	forretning (shop/business)	menneske (person/human being)	3.928	0.155	8844	116
10	kælk (toboggan/sledge)	slæde (sledge)	10.458	0.022	1118	196	menneske (person/human being)	standse (to stop)	3.911	0.164	5739	48
11	oldefar (greatgrandfather)	olde(mor) (greatgrandmother)	10.016	0.030	72	34	holde øje med (to watch)	standse (to stop)	3.906	0.171	3647	31
12	forretning (shop/business)	menneske (person/human being)	9.849	0.023	1138	211	forretning (shop/business)	holde øje med (to watch)	3.904	0.161	6793	67
13	slæde (sledge)	rar (pleasant/nice)	9.791	0.021	2523	478	holde øje med (to watch)	kravlegård (playpen)	3.885	0.156	9321	131
14	kælk (toboggan/sledge)	menneske (person/human being)	9.355	0.019	5554	1147	forretning (shop/business)	gyngestol (rocking chair)	3.702	0.163	9555	175
15	hjort (deer/stag)	hvalp (puppy)	9.303	0.026	639	132	hjort (deer/stag)	pony	3.688	0.162	10314	203
16	menneske (person/human being)	mørk (dark)	9.201	0.022	2967	576	pony	kravlegård (playpen)	3.672	0.164	9798	189
17	menneske (person/human being)	kalkun (turkey)	9.192	0.025	1090	217	slæde (sledge)	rar (pleasant/nice)	3.660	0.143	20334	1239
18	standse (to stop)	kravlegård (playpen)	9.135	0.023	2001	399	slæde (sledge)	slem (bad)	3.649	0.144	20379	1262
19	menneske (person/human being)	kravlegård (playpen)	8.953	0.023	2778	553	slæde (sledge)	kravlegård (playpen)	3.637	0.137	24504	2183
20	menneske (person/human being)	holde øje med (to watch)	8.931	0.021	4244	887	hjort (deer/stag)	kalkun (turkey)	3.623	0.161	12015	303
21	standse (to stop)	æske (small box)	8.919	0.023	2772	554	menneske (person/human being)	æske (small box)	3.614	0.161	12088	315
22	forretning (shop/business)	kravlegård (playpen)	8.894	0.029	341	79	standse (to stop)	æske (small box)	3.611	0.176	6031	89
23	holde øje med (to watch)	pæn (nice)	8.890	0.028	588	122	forretning (shop/business)	bagved (behind)	3.605	0.176	6127	93
24	sætte (to sit)	vise (to show)	8.869	0.043	20	11	slæde (sledge)	æske (small box)	3.592	0.146	20628	1390
25	holde øje med (to watch)	kravlegård (playpen)	8.810	0.026	1206	247	forretning (shop/business)	æske (small box)	3.545	0.169	10080	233
26	slutte (to end)	pæn (nice)	8.777	0.030	287	69	bagved (behind)	kravlegård (playpen)	3.514	0.173	9118	208
27	menneske (person/human being)	trække (to pull)	8.697	0.029	462	105	tage (to take)	tørre af (to dry)	3.504	0.220	239	4
28	pæn (nice)	ren (clean)	8.678	0.034	97	41	oldefar (greatgrandfather)	olde(mor) (greatgrandmother)	3.497	0.138	26304	3148
29	hans (his)	bagved (behind)	8.436	0.028	987	215	gyngestol (rocking chair)	kravlegård (playpen)	3.457	0.159	16705	926
30	forretning (shop/business)	gyngestol (rocking chair)	8.415	0.029	721	158	kirke (church)	menneske (person/human being)	3.447	0.165	14074	600

Table 8. Illustration of the hazard approach, where the comparison was limited to 1743 pairs. Ranking is done after the lower bound of the confidence interval CI. Additional ranking R after θ and $ratio_2$ is presented as well as the number of children, where the absolute difference could be observed.

Rank	Pair		CI lower	θ	CI upper	R θ	R $ratio_2$	# of observed differences
1	bedstefar (grandfather)	bedste(mor) (grandmother)	8.142	8.499	8.855	3	1	66
2	hjort (deer/stag)	kravlegård (playpen)	7.739	9.274	10.886	1	32	28
3	slæde (sledge)	faster (paternal aunt)	7.152	9.022	11.091	2	39	27
4	pæn (nice)	rar (pleasant/nice)	6.771	7.576	8.427	11	36	39
5	forretning (shop/business)	kravlegård (playpen)	6.727	8.172	9.729	4	3	40
6	slæde (sledge)	forretning (shop/business)	6.488	7.974	9.613	5	6	30
7	kirke (church)	menneske (person/human being)	6.361	7.370	8.465	14	30	34
8	slæde (sledge)	menneske (person/human being)	6.355	7.253	8.234	18	2	27
9	forretning (shop/business)	hjort (deer/stag)	6.325	7.796	9.344	7	187	28
10	menneske (person/human being)	oldefar (greatgrandfather)	6.298	7.784	9.443	8	510	27
11	kælk (toboggan/sledge)	faster (paternal aunt)	6.284	7.727	9.267	9	562	34
12	menneske (person/human being)	kravlegård (playpen)	6.242	6.881	7.563	38	5	29
13	kirke (church)	kravlegård (playpen)	6.082	7.132	8.272	23	64	37
14	forretning (shop/business)	kirke (church)	6.072	6.830	7.624	42	41	37
15	gyngestol (rocking chair)	kravlegård (playpen)	6.000	7.006	8.106	29	29	33
16	affald (garbage)	kravlegård (playpen)	5.921	7.239	8.688	20	178	35
17	slæde (sledge)	rar (pleasant/nice)	5.902	7.816	10.068	6	17	25
18	menneske (person/human being)	onkel (uncle)	5.894	7.394	9.107	12	468	29
19	forretning (shop/business)	menneske (person/human being)	5.822	6.760	7.786	48	9	35
20	kælk (toboggan/sledge)	kravlegård (playpen)	5.789	7.296	8.987	17	160	29
21	forretning (shop/business)	holde œje med (to watch)	5.777	7.578	9.604	10	12	31
22	menneske (person/human being)	holde œje med (to watch)	5.759	6.568	7.452	69	4	28
23	elske (to love)	kravlegård (playpen)	5.753	6.970	8.295	33	423	35
24	kælk (toboggan/sledge)	forretning (shop/business)	5.733	6.988	8.376	30	322	31
25	faster (paternal aunt)	tœrre af (to dry)	5.730	7.115	8.624	24	632	33
26	made (to feed)	standse (to stop)	5.729	6.422	7.161	92	593	29
27	oldefar (greatgrandfather)	olde(mor) (greatgrandmother)	5.720	6.532	7.385	76	28	46
28	hjort (deer/stag)	pony	5.689	6.714	7.824	55	15	35
29	forsigtig (careful)	affald (garbage)	5.675	6.937	8.288	35	729	38
30	rar (pleasant/nice)	forsigtig (careful)	5.651	7.030	8.543	27	1302	32

Table 9. Taking gender and lexicon size into account: Illustration of the hazard approach, where the comparison was limited to 1743 pairs (cf. Table 8). Ranking is done after θ. Additional ranking R after θ from Table 8 is presented.

	gender				lexicon size			
Rank	Pair		θ	R θ	Pair		θ	R θ
1	bedstefar (grandfather)	bedste(mor) (grandmother)	4.837	3	sige (to say)	hvornår (when?)	1.192	592
2	faster (paternal aunt)	tallerken (plate)	4.719	198	forretning (shop/business)	skål (bowl)	1.137	442
3	moster (maternal aunt)	olde(mor) (greatgrandmother)	4.638	77	god (good)	slem (bad)	1.122	1671
4	bedste(mor) (grandmother)	falde (to fall)	4.509	904	slæde (sledge)	shorts	1.121	418
5	bedstefar (grandfather)	tegne (to draw)	4.474	1082	slæde (sledge)	krus (cup)	1.093	759
6	faster (paternal aunt)	tå (toe)	4.453	447	slæde (sledge)	sige (to say)	1.090	207
7	faster (paternal aunt)	barnevogn (baby carriage)	4.418	467	forretning (shop/business)	shorts	1.085	981
8	faster (paternal aunt)	bide (to bite)	4.393	788	farfar (paternal grandfather)	olde(mor) (greatgrandmother)	1.082	726
9	faster (paternal aunt)	køre (to drive)	4.378	696	skål (bowl)	gyngestol (rocking chair)	1.067	1672
10	slæde (sledge)	sokker (socks)	4.356	257	slæde (sledge)	glad (happy)	1.065	473
11	faster (paternal aunt)	moster (maternal aunt)	4.349	50	bedste(mor) (grandmother)	kalkun (turkey)	1.062	218
12	hjort (deer/stag)	skål (bowl)	4.349	154	faster (paternal aunt)	tå (toe)	1.044	447
13	morfar (maternal grandfather)	moster (maternal aunt)	4.304	976	kælk (toboggan/sledge)	kravlegård (playpen)	1.036	17
14	bedste(mor) (grandmother)	tørstig (thirsty)	4.259	1063	menneske (person/human being)	shorts	1.036	348
15	faster (paternal aunt)	pasta	4.259	327	få (to get)	hvornår (when?)	1.030	885
16	kælk (toboggan/sledge)	faster (paternal aunt)	4.237	9	smide (to throw)	hjort (deer/stag)	1.027	619
17	moster (maternal aunt)	nattøj (pyjamas)	4.236	1019	faster (paternal aunt)	pasta	1.027	327
18	babysitters navn (babysitter's name)	bedste(mor) (grandmother)	4.226	208	forretning (shop/business)	dette / det her (this/that)	1.025	138
19	faster (paternal aunt)	potte (potty)	4.225	409	bedstefar (grandfather)	tegne (to draw)	1.024	1082
20	slæde (sledge)	faster (paternal aunt)	4.204	2	shorts	kravlegård (playpen)	1.004	923
21	oldefar (greatgrandfather)	olde(mor) (greatgrandmother)	4.201	76	rydde op (to clean up)	kravlegård (playpen)	1.004	795
22	faster (paternal aunt)	kjole (dress)	4.148	384	slæde (sledge)	stranden (the beach)	1.001	727
23	babysitters navn (babysitter's name)	bedstefar (grandfather)	4.140	151	sige (to say)	kravlegård (playpen)	1.000	537
24	faster (paternal aunt)	give (to give)	4.134	130	faster (paternal aunt)	lampe (lamp)	1.000	733
25	faster (paternal aunt)	olde(mor) (greatgrandmother)	4.133	103	faster (paternal aunt)	sår (wound)	0.987	520
26	faster (paternal aunt)	sige (to say)	4.100	1570	give (to give)	hans (his)	0.981	563
27	faster (paternal aunt)	køkken (kitchen)	4.093	952	bedste(mor) (grandmother)	elske (to love)	0.973	427
28	faster (paternal aunt)	denne / den her (this/that)	4.052	572	slæde (sledge)	give (to give)	0.969	80
29	slæde (sledge)	give (to give)	4.044	80	babysitters navn (babysitter's name)	bedste(mor) (grandmother)	0.932	208
30	oldefar (greatgrandfather)	skål (bowl)	4.032	445	onkel (uncle)	badeværelse (bathroom)	0.925	994

8 Summary and outlook

In this paper, we discussed the possibility of measuring and estimating closeness for a pair of event times based on the distribution of their absolute difference, restricted to $t_* \leq \tilde{T}_{min} \leq t^*$. While this necessary restriction might influence absolute results, it does not hinder the comparison of many pairs, which was the main aim in the motivating example. We also presented several concrete approaches, absolute as well as relative measures, and their implementation proved in principle easily feasible based on standard statistical software.

Applying two simple, exemplary absolute measures in the example data yielded sensible and interpretable results: words which are likely to be learnt and/or reported closely together turned up on the list of the 30 closest pairs. Not surprisingly, the choice of x in the absolute measures m_x showed a large effect on the results: candidates for absolute measures have to be chosen very carefully.

Setting the absolute measures into perspective by comparing it to their equivalent under a reference distribution turned out to be feasible as well, yielding interesting results when applied in the data. However, the choice of how to compare measures pointed to some methodological problems: the comparison based on the difference was easily possible but did not entail the intended re-assessment of the word pairs, whereas comparison based on the ratio seemed more meaningful but involved (possible) division by very small numbers, even without considering conditioning on $\tilde{T}_{min}$. The hazard approach presented a third alternative to a relative measure, pointing out similar close word pairs as the ratio approach. Here, again, the choice of x, up to which the hazard ratios are assumed to be constant, seems crucial.

We also discussed the possibility to take covariates into account to improve comparison. While this seems possible in theory, practical implementation relies on a well-working relative measure, which is both meaningful and practically feasible. Therefore, analysis results were but an illustration of shortcomings of our candidates for the relative measures.

To summarize, the concept of closeness, as we approached it here, seems to be working well, yet the concrete implementation needs further work and refinement. To name just a few examples, the behaviour of the presented closeness measures should be properly investigated, in a broad range of simulated and real data settings. Other, potentially more suitable candidates for measures might be identified and studied. Further, the appropriateness of the simple linear Cox model for estimation should be addressed. It also remains to be studied whether other approaches, say, transformation of the time axis instead of taking into account a covariate, would be feasible etc. For the effort to be worthwhile, it also needs to be shown, that closeness is useful in other contexts as well.

But despite all these un-addressed issues, we seem nevertheless on the right track to be able to answer the motivating research question in our example: yes, there are words which are learnt closer together than expected under chance condition. And to clearly identify the influential linguistic factors by more extensive analyses, would be the obvious next step.

9 Acknowledgements

We would like to thank everybody involved in the Danish CDI project, especially all families who participated in the labour-intensive longitudinal study. The data collection was initiated by the Odense Project of Language Acquisition, supported by the Danish Research Council for the Humanities.

10 References

C. Andersen, W, Vach, S. Wehberg and D. Bleses, "Conduct of the Danish CDI Studies", *Working papers, Center for Child Language e-prints* vol. 1, 2006 (retrieved from Center for Child Language website, at the University of Southern Denmark, Odense).

J.M. Bland and D.G. Altman,"Statistical method for assessing agreement between two methods of clinical measurement", *The Lancet* vol. i pp. 307-310, 1986.

D. Bleses, W. Vach, S. Wehberg, K. Faber and T.O. Madsen, *Tidlig kommunikativ udvikling. Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI-forældrerapportundersøgelser af danske normalthørende og hørehæmmede børn* [Early communicative development. An instrument to describe language acquisition of Danish normal hearing and hearing impaired children based on CDI parental-report studies], University Press of Southern Denmark, Odense, in press.

D. Bleses, W. Vach, M. Slott, S. Wehberg, P. Thomsen, T.O. Madsen and H. Basbøll, "The Danish Communicative Development Inventories: validity and main developmental trends", submitted to *Journal of Child Language.*

J. Fan, L. Hsu and R. Prentice, "Dependence estimation over a finite bivariate failure time region", *Lifetime Data Analysis* vol. 6 pp. 343-355, 2000.

L. Fenson, P. Dale, J.S. Reznick, D. Thal, E. Bates, J. Hartung, S. Pethik and J. Reilly, *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*, Singular Publishing Group, San Diego, 1993.

L. Fenson, P. Dale, J.S. Reznick, E. Bates, D. Thal and S. Pethik, "Variability in Early Communicative Development", *Monographs of the Society for Research in Child Development* vol. 59(5), Serial no. 242, 1994.

L. Fenson, V. Marchman, D. Thal, P. Dale, J.S. Reznick and E. Bates, *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual*, 2nd edition, Paul H. Brookes Publishing Co, Baltimore, 2007.

D. Oakes, "Bivariate survival models induced by frailties", *Journal of the American Statistical Association* vol. 84 pp. 487-493, 1989.

S. Wehberg, W. Vach, D. Bleses, P. Thomsen, T.O. Madsen and H. Basbøll, "Danish children's first word: Analysing longitudinal data based on monthly CDI parental reports", to appear in *First Language*, 2007.

DANSK RESUMÉ

Indledning[1]

Børns sprogtilegnelse, som er "afkommet" af hovedsagelig lingvistik og udviklingspsykologi, er et ret nyt forskningsområde. Dets "fødsel" kan føres tilbage til 1974, da *The Journal of Child Language (JCL)* blev grundlagt. Seks år senere blev *First Language (FL)* grundlagt. Begge tidsskrifter betragtes som toneangivende indenfor deres område og bringer "articles on all aspects of the scientific study of language behaviour in children, the principles which underlie it, and the theories which may account for it[2]" (JCL homepage[3]) og "original research, theoretical articles, review articles and book reviews in all areas of first language acquisition[4]" (FL homepage[5]). En hurtig søgning på "Danish" gav 11 resultater i JCL og 9 resultater i FL, hvor næsten alle har forbindelse til forskningen udført af Plunkett og kolleger (f.eks. Plunkett 1984; Plunkett 1986; Plunkett & Strömqvist 1992; se også det danske korpus i CHILDES udført af Plunkett[6]): den internationalt offentliggjorte forskning om dansk børnesprogtilegnelse i det mindste indenfor de seneste 30 år er begrænset (se Bleses m.fl. *submitted 2*).

Odense Projektet i Sprogtilegnelse (1998-2001) blev grundlagt af en tværfaglig forskningsgruppe for at fremme forskning i første sprogtilegnelse på dansk (Basbøll m.fl. 2002). En del af dette projekt var, at adaptere det amerikanske instrument *MacArthur-Bates Communicative Development Inventories (CDI)* (Fenson m.fl. 1993; Fenson m.fl. 1994; Fenson m.fl. 2007), som er en meget brugt forældrerapport i form af en checkliste, til dansk. Dette instrument er opdelt i et skema "Words and gestures" (også betegnet som *Infant del*) rettet mod

[1] Bemærk at i dette sammendrag vil citater fra resten af denne Ph.D. afhandling, inklusive artikler i kapitel 3 til 5, ikke være eksplicit afmærket.

[2] Oversættes til "artikler om alle aspekter af det videnskabelige studie af sproglig adfærd hos børn, principperne, som ligger til grund for den og teorierne, som kan forklare den".

[3] *http://journals.cambridge.org/action/displayJournal?jid=JCL.*

[4] Oversættes til: "original forskning, teoretiske artikler, anmeldelser og boganmeldelser indenfor alle områder af første sprogtilegnelse".

[5] *http://www.sagepub.com/journalsProdDesc.nav?prod=Journal201667.*

[6] *http://childes.psy.cmu.edu.*

aldersgruppen 8-16 måneder og et skema "Words and gestures" (*Toddler del*) i aldersgruppen 16-30 måneder. CDI'er var udformet til at bedømme børns tidlige kommunikative færdigheder på forskellige områder ved f.eks. at spørge forældre om deres barn bruger visse gestikulationer (Infant del), forstår og/eller siger visse almindelige ord (begge dele), eller er begyndt at bruge mere komplekse sætningskonstruktioner (Toddler del). Med dette værktøj som udgangspunkt søgte forskerne at beskrive det (gennemsnitlige) forløb af den tidlige danske sprogtilegnelse og at sammenligne danske børns sproglige udvikling på tværs af regioner, kulturer og forskellige sprog (se Andersen m.fl. 2006).

Dataindsamlingen, baseret på det danske CDI instrument, havde to formål: en storstilet tværsnitsundersøgelse, som omfattede 6112 danske børn, blev udført for at belyse udviklingen af ordforråd, og give stof til analyser af, hvilken indflydelse faktorer som køn og/eller forældrenes uddannelse har på den tidlige sprogtilegnelse (Bleses m.fl. *under udgivelse*; Bleses m.fl. *submitted 1*; Bleses m.fl. *submitted 2*). Dette blev ledsaget af en længde-undersøgelse for at blive i stand til at studere børns individuelle udvikling over tid. Med omkring 180 deltagere er det danske longitudinale CDI studie en storstilet undersøgelse og repræsenterer, så vidt jeg ved, den største datakilde blandt de CDI-baserede longitudinale studier indtil videre: det er et unikt datasæt, som tilbyder næsten utallige muligheder for analyse.

Et tværfagligt Ph.D. projekt, som resulterede i nærværende Ph.D. afhandling, blev iværksat i 2003 for at hjælpe til med at analysere de indsamlede data, og dets bredt formulerede formål var "at udvikle og anvende statistiske metoder som er passende til at afprøve hypoteser om sprogtilegnelse blandt danske børn".

Data

Hovedformålet med de tværsnitlige data var at etablere normer og sammenligne danske børns gennemsnitlige tidlige sprogudvikling tværkulturelt og tværlingvistisk baseret på målinger, som var prædefineret af andre offentligjorte resultater, d.v.s. hovedsagligt "sum scores". Fokus her ligger at andet sted: kilden til al forskning, som præsenteres i det følgende er data fra den danske CDI longitudinale undersøgelse. Med udgangspunkt i det danske CDI værktøj blev 183 danske børn fulgt månedligt fra 8 måneders alderen til 30 måneders alderen, idet den danske Infant del af CDI ("Ord og gestikulation") blev benyttet fra 8 til 15 måneders alderen og den danske Toddler del ("Ord og sætninger") fra 16. måned til undersøgelsens afslutning. Analyserne fokuserer hovedsagligt på de dele af CDI checklisten, som har med ordforråd at gøre, nærmere betegnet de 410 ord (begreber), som findes i både Infant og Toddler delen. Dataenes longitudinale struktur giver os mulighed for at følge optræden af "første ord" hos et barn, og derfor bliver det muligt at analysere tilegnelsestidspunktet (hvis man stiller tilegnelsen på samme trin som rapportering for første gang). Den egentlige definition af tilegnelsestidspunktet for et specifikt begreb, d.v.s. tidspunktet hvor det første ord bliver nævnt (krydset af som "sagt og forstået" af forælderen), kan udføres på mindst to forskellige tidsskalaer: den kronologiske aldersskala som den måned, hvor ordet først optræder (f.eks. i måned 12) eller skalaen for den individuelle ordforrådsstørrelse, hvor den rangerer i et barns tilegnelsessekvens (f.eks. som ord nr. 56).

Formål

Da dette er et tværdisciplinært projekt, som både kombinerer børnesprogstilegnelse og statistik, havde denne afhandling to formål. Det første formål, som repræsenterer sprogtilegnelse, var at få indsigt i og dokumentere resultaterne af danske børns første brug af sprog. Ved at opdele dette generelle formål i flere "mindre" formål, begyndte vi med at beskrive danske børns første ord og sammenligne dem tværlingvistiskt, derefter studerede vi både gruppe og individuelle variationer: afhænger disse ord af faktorer som f.eks. køn, og er der børn som udviser (individuelle) præferencer for visse ordklasser? Det sidste sprog-relaterede forskningsspørgsmål omhandlede forholdet mellem ordpar: kan vi identificere par, der er tilegnet tættere sammen end forventet og deler disse nogle genkendelige lingvistiske egenskaber, som f.eks. at de rimer? Det andet formål var af en mere metodologisk, statistisk art. Vi ønskede at udforske CDIs potentiale udover beregningen af "sum scores": vil analyse af enkelte begreber på ordforrådschecklisten, især med time-to-event metoder på ordforrådsstørrelseskalaen vise sig at bære frugt? En yderligere statistisk øvelse var udviklingen af en målemetode til formelt at kvantificere "closeness" for et ordpar.

Resultater og diskussion

For det første var vi i stand til at reproducere ”typiske” resultater af første ord baseret på f.eks. deres semantiske-pragmatiske indhold og deres lydstruktur. Set i lyset af undersøgelsesresultater, der indikerer at danske børn typisk er 2-3 måneder bagefter deres amerikanske jævnalderende m.h.t. tidlig forståelse og produktion af ord (Bleses m.fl. *submitted 2*), var dette beroligende, eftersom vi intet fandt, der støtter hypotesen om at danske børn følger en sprogtilegnelsesrækkefølge eller et -mønster, som er forskelligt fra deres amerikanske eller italienske jævnaldrende. Vi har derimod identificeret nogle slående tværlingvistiske forskelle på enkeltordsniveauet, hvilket giver grund til at revurdere hypoteser om arten og optræden af første ord i almindelighed. Vi var også i stand til at finde, ligeledes på enkeltordsniveauet, forskelle i den gennemsnitlige tilegnelsestid (målt på ordforrådsstørrelsesskalaen), der henfører til grupper, som er defineret af køn og søskende status, og vi formulerede nogle ad-hoc kategorier, som redegør for disse forskelle. Med hensyn til individualiteten af børns sammensætning af leksikon, opdagede vi, at visse børn udviste præferencer, som bestod over tid for nogle ordklasser på dette tidlige stadie. Disse individuelle forskelle blandt børn kan måske sættes i forbindelse med distinktioner som *referentiel* mod *ekspressiv* (se Shore 1995). I en undersøgelse af det gensidige forhold mellem tilegnelsestid og specifikke CDI ordpar, opnåede vi resultater, der tyder på, at ord, der deler semantiske og pragmatiske træk, som f.eks. kropsdele, (zoo) dyr eller betegnelser for slægtninge, havde en tendens til at blive rapporteret tidsmæssigt tæt sammen. Ved at anvende en mere eksakt, relativ måling fandt vi også nogle spændende mønstre (mellem de pågældende ordpar). Disse (mønstre) er imidlertid ikke så knyttede til ordenes indhold, som resultaterne fra den mere simple metoder tyder på. Blandt andet virker hypotesen om at ”ordets længde” har en indflydelse som forbindende faktor plausibel og bør undersøges nærmere.

Med hensyn til det andet formål understøttede alle analyseresultaterne CDIs nytte på trods af den megen diskussion af metodens begrænsninger (f.eks. Pine 1992). I særdeleshed kunne CDI målingen fastholde forskelle på enkeltordsniveau, ligesom den kunne fange individuel variation over tid. Med longitudionale data som vore var anvendelsen af time-to-event metoder mulig og lovende. For at give en første illustration, virkede Cox proportional hazards modellen (på ordforrådsstørrelseskalaen) fint til at identificere enkelte begreber, som varierede i tilegnelsestid blandt grupper; vi forslog yderligere en måling som "closeness" for et enkelt ordpar baseret på den absolutte difference på de tilsvarende tilegnelsestider, i forhold til den forventede difference under (konditionel) uafhængighed. Denne måling kan også vise sig at være nyttig i andre forbindelser, hvor det er af interesse at måle "closeness" af to "event" tider.

Konklusion

Analysen af data fra den danske longitudinale CDI undersøgelse, som forgik fra to vinkler – sprogtilegnelse og statistik – viste sig at være inspirerende for begge parter. Indenfor sprogtilegnelse bliver der sædvanligvis brugt mange anstrengelser, tid og penge på data indsamling. At udvikle og anvende (statistiske) metoder, som kan hjælpe med til at udtrække så mange informationer som muligt, synes at være anstrengelserne værd, og det har vist sig at bære frugt: vore analyser gav interessante og meningsfyldte resultater. Med hensyn til statistik er statistisk videnskab forholdsvis meningsløs uden data. Den danske longitudinale CDI undersøgelse er på mange måder et unikt datasæt, og det har ikke kun været spændende i sig selv at finde passende måder at løse forskningsspørgsmål på, men kan også resultere i statistiske metoder, som kan anvendes indenfor andre områder. Men når jeg ser på Ph.D. projektets bredt formulerede emne ovenfor, må jeg konkludere, at selv om det var morsomt, er jeg langt fra at være færdig.

Opbygning af denne afhandling

Denne afhandling består af fem kapitler. Kapitel 1 giver et udvidet sammendrag af afhandlingen, som sætter resultaterne af de efterfølgende enkelte papers (kapitler) ind i et bredere perspektiv. Kapitel 2 begynder med at præsentere data fra den danske longitudinale CDI undersøgelse i detaljer, og fremhæver derefter forskellige aspekter i forbindelse med den danske CDIs validitet og pålidelighed. Næst i kapitlet præsenteres test-retest korrelationer og prædiktive værdier, som karakteristika af den danske CDI. Kapitel 3 består af det første paper *Danish children's first words – Analysing longitudinal data based on monthly CDI parental reports* (Paper 1) og kapitel 4 indeholder det ledsagende paper *Girls talk about dolls and boys about* *cars? Analyses of group and individual variation in Danish children's first words* (Paper 2) sammen med en tilføjelse, der indeholder nogle tekniske og teoretiske overvejelser angående analyserne, som er beskrevet i Paper 2. Den tredje artikel *Measuring closeness of two event times* findes i kapitel 5. En sammenfatning på dansk udgør det sidste afsnit før Appendiks, som indeholder skanninger af de danske CDI skemaer.

Bibliografi

Andersen, C., Vach, W., Wehberg, S. & Bleses, D. (2006): Conduct of the Danish CDI Studies. *Working papers, Center for Child Language e-prints*, no. 1. Retrievable from the website of the Center for Child Language, University of Southern Denmark, Odense.

Basbøll, H., Bleses, D., Cadierno, T., Jensen, A., Ladegaard, H.J., Madsen, T.O., Millar, S., Sinha, C. & Thomsen P. (2002): The Odense Language Acquisition Project. *Child Language Bulletin 22*, 11-12.

Bleses, D., Vach, W., Wehberg, S., Faber, K. & Madsen, T.O. (in press): *Tidlig kommunikativ udvikling. Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI-forældrerapportundersøgelser af danske normalthørende og hørehæmmede børn* [Early communicative development. An instrument to describe language acquisition of Danish normal hearing and hearing impaired children based on CDI parental-report studies]. University Press of Southern Denmark, Odense.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 1): The Danish Communicative Development Inventories: validity and main developmental trends. Submitted to *Journal of Child Language*.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T.O. & Basbøll, H. (submitted 2): Building the case for a Danish delay: Early vocabulary development in Danish children in a cross-linguistic CDI-based comparison. Submitted to *Journal of Child Language*.

Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethik, S. & Reilly, J. (1993): *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Singular Publishing Group, San Diego.

Fenson, L., Dale, P., Reznick, J.S., Bates, E., Thal, D. & Pethik, S. (1994): Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development 59(5)*, Serial no. 242.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J.S. & Bates, E. (2007): *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* Second edition. Paul H. Brookes Publishing Co, Baltimore.

Pine, J. (1992): How referential are "referential" children? Relationships between maternal-report and observational measures of vocabulary composition and usage. *Journal of Child Language 19*, 75-86.

Plunkett, K. (1984): Projekt barnesprog: some preliminary analysis and discussion. *First Language 5*, 219-220.

Plunkett, K. (1986): Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology 27*, 64-73.

Plunkett, K. & Strömqvist, S. (1992): The acquisition of Scandinavian languages. In D. Slobin (ed.): *The crosslinguistic study of language acquisition*, Vol. 3. Lawrence Erlbaum Associates. Chapter 7, 457-556.

Shore, C. (1995): *Individual Differences in Language Development.* Sage Series on Individual Differences and Development. Sage Publications.

Appendix

Appendix A. Sample scan of the Danish CDI Infant part form *Ord og gestikulation*

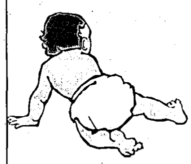
On the following pages, a sample scan of the Danish CDI Infant part form *Ord og gestikulation* is provided for illustration purposes only (version from February 29th, 2000). Under no circumstances it is permitted to use this sample as a basis for study conduct or other research purposes. See Bleses et al. (*in press*) for purchasing of the Danish CDI forms (University Press of Southern Denmark, Campusvej 55, 5230 Odense M, Tlf. +45 66 15 79 99, www.universitypress.dk).

"We thank Larry Fenson and colleagues for permission to adapt the MacArthur-Bates Communicative Development Inventory to Danish. The adaptation into the Danish CDI was part of a larger project on language acquisition, *The Odense Language Acquisition-project* (1998-2001), funded by The Danish Research Council for the Humanities and University of Southern Denmark (The adaptation of *CDI: Words and Gestures* was done in collaboration with the Paediatric Nutrition Group, Department of Human Nutrition, The Royal Veterinary and Agricultural University, Copenhagen, and we are grateful for their contribution)." (Bleses et al., *submitted 2*, Acknowledgements)

For references, see chapter 1.

Barnets navn:__________________ **Fødselsdag:**________ **Dags dato:**________

MacArthur CDI:
Ord og gestikulation



The MacArthur Communicative Development Inventory:
Words and Gestures

For Information/copies, contact The Developmental Psychology
Lab, San Diego State University, San Diego, CA 92182

Tilpasset til dansk af

Odense-Projektet i Sprogtilegnelse
Institut for Sprog og Kommunikation,
Syddansk Universitet, Odense Universitet
Kontaktadresse: Campusvej 55, 5230 Odense M
&
Børneernæringsgruppen
Forskningsinst. for Human Ernæring, Kbh.
1999

A. FØRSTE TEGN PÅ FORSTÅELSE

Før børn begynder at tale, viser de tegn på, at de forstår talen; ofte reagerer de på velkendte ord og vendinger. Nedenfor er nogle typiske eksempler. Reagerer dit barn på nogle af disse?

	ja	nej
1. Reagerer på eget navn	o	o
2. Reagerer på 'nej' (ved at stoppe med den handling han/hun er i gang med - i det mindste et øjeblik)	o	o
3. Reagerer på 'der er mor/far' ved at kigge efter dem	o	o

B. (Mere eller mindre faste) VENDINGER (26)

Marker venligst de vendinger nedenfor, som du tror, at dit barn forstår

	forstår		forstår		forstår
det må du ikke	O	ikke røre/pille	O	skal vi gå?	O
det vil jeg ikke ha'	O	kast bolden	O	skifte (ble)	O
er du sulten?	O	klap (i hænderne)	O	spyt den/det ud	O
er du træt/søvnig?	O	kom (herhen/nu)	O	stor dreng/pige	O
giv den/det til mor	O	nu skal du putte	O	stå/lig stille	O
giv mig et knus	O	pas på/forsigtig	O	ti stille	O
giv mig et kys	O	rejs dig op	O	vil du have mere?	O
gå en tur	O	se her	O	åbn munden	O
hent ____	O	sid ned	O		

C. DEN FØRSTE TALE

		aldrig	nogle gange	ofte
1.	Nogle børn kan lide at 'snakke efter' eller at efterligne (imitere) noget, de lige har hørt (også nye ord, de er ved at lære og/eller dele af sætninger). Fx kan de gentage "arbejde nu" lige efter, at mor har sagt "Mor skal på arbejde nu". Hvor ofte imiterer dit barn?	O	O	O
2.	Nogle børn kan lide at opremse navne og betegnelser på ting, som om de er stolte over at kende ordene og ønsker at vise dette. Hvor ofte gør dit barn det?	O	O	O

D. ORDFORRÅD - CHECKLISTE

Det følgende er en liste over typiske ord i børns ordforråd. De ord som dit barn *forstår*, men endnu ikke siger, skal markeres i første række (*forstår*). De ord som dit barn både *forstår* og *siger* skal markeres i anden række (*forstår* og *siger*). Selvom dit barn har en [illegible] udtale af et ord (fx 'gugge' for 'dukke', 'dol' for 'stol' eller 'detti' for 'spaghetti'), bedes du alligevel markere ordet. Husk at dette er et 'katalog' over ord, der kan findes hos mange forskellige børn. Vær ikke bekymret, hvis dit barn kun kender nogle få ord lige nu.

1. Lydeffekter og dyrelyde (11)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
av	○	○	mjav	○	○	rap	○	○
fut	○	○	muh	○	○	vov	○	○
grrrr	○	○	mæh	○	○	årnnn (bil-lyd)	○	○
kykliky	○	○	mm mm (lækkert)	○	○			

2. Dyrenavne (virkelige eller legetøj) (36)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
abe	○	○	får	○	○	ko	○	○
and	○	○	giraf	○	○	kylling	○	○
bamse	○	○	gris	○	○	lam	○	○
bi	○	○	gås	○	○	løve	○	○
bjørn	○	○	hest	○	○	mus	○	○
dyr	○	○	hjort	○	○	pingvin	○	○
egern	○	○	hvalp	○	○	pony	○	○
elefant	○	○	hund	○	○	skildpadde	○	○
fisk	○	○	kalkun	○	○	sommerfugl	○	○
flue	○	○	kanin	○	○	tiger	○	○
frø	○	○	kat	○	○	ugle	○	○
fugl	○	○	killing	○	○	æsel	○	○

3. Transportmidler (virkelige eller legetøj) (10)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
barnevogn	○	○	cykel	○	○	motorcykel	○	○
bil	○	○	flyvemaskine	○	○	tog	○	○
brandbil	○	○	klapvogn	○	○			
bus	○	○	lastbil	○	○			

MacArthur CDI. 3 *Ord og gestikulation* – dansk version

4. Legetøj (8)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
ballon	O	O	bold	O	O	legetøj	O	O
blyant	O	O	dukke	O	O	sæbebobler	O	O
bog	O	O	klods	O	O			

5. Mad og drikke (28)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
appelsin	O	O	kaffe	O	O	rosiner	O	O
banan	O	O	kage(r)	O	O	slik	O	O
boller	O	O	kiks	O	O	smør	O	O
brød	O	O	kylling	O	O	spaghetti	O	O
cornflakes	O	O	kød	O	O	vand	O	O
fisk	O	O	mad	O	O	æble	O	O
gulerødder	O	O	mælk	O	O	æg	O	O
havregryn	O	O	ost	O	O	ærter	O	O
is	O	O	pasta	O	O			
juice	O	O	pizza	O	O			

6. Tøj (21)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
ble	O	O	jakke	O	O	sko	O	O
bluse	O	O	kjole	O	O	sokker	O	O
t-shirt	O	O	knap	O	O	strømper	O	O
bukser	O	O	lynlås	O	O	støvler	O	O
frakke	O	O	nattøj	O	O	sut	O	O
hagesmæk	O	O	halskæde	O	O	sweater	O	O
hue	O	O	shorts	O	O	trøje	O	O

7. Legemsdele (20)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
ansigt	O	O	kind	O	O	tissemand	O	O
arm	O	O	knæ	O	O	tunge	O	O
ben	O	O	mund	O	O	tænder	O	O
finger	O	O	navle	O	O	tå	O	O
hoved	O	O	næse	O	O	øje	O	O
hånd	O	O	sår	O	O	øre	O	O
hår	O	O	tissekone	O	O			

8. Små husholdningsting (39)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
affald	O	O	kop	O	O	saks	O	O
billede	O	O	kost	O	O	ske	O	O
briller	O	O	krus	O	O	skrald	O	O
børste	O	O	lampe	O	O	skål	O	O
dyne	O	O	lys	O	O	småpenge	O	O
flaske	O	O	medicin	O	O	støvsuger	O	O
gaffel	O	O	nøgler	O	O	sæbe	O	O
glas	O	O	papir	O	O	tallerken	O	O
hammer	O	O	penge	O	O	tandbørste	O	O
håndklæde	O	O	plante	O	O	telefon	O	O
kam	O	O	pude	O	O	tæppe	O	O
kasse	O	O	pung	O	O	ur	O	O
klokke	O	O	radio	O	O	æske	O	O

9. Møbler og rum (24)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
badekar	O	O	kravlegård	O	O	soveværelse	O	O
badeværelse	O	O	køkken	O	O	stol	O	O
bord	O	O	køleskab	O	O	stue	O	O

dør	O	O	ovn	O	O	trapper	O	O
garage	O	O	potte	O	O	tv/fjernsyn	O	O
gyngestol	O	O	seng	O	O	vask	O	O
høj stol	O	O	skuffe	O	O	vindue	O	O
komfur	O	O	sofa	O	O	vugge		

10. Udendørsting (14)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
blomst	O	O	regn	O	O	sten	O	O
gynge	O	O	skovl	O	O	stjerne	O	O
himmel	O	O	slæde	O	O	træ	O	O
kælk	O	O	sne	O	O	vand	O	O
måne	O	O	sol	O	O			

11. Steder (man kan tage hen til) (14)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
arbejde	O	O	hjem	O	O	stranden	O	O
butik	O	O	hus	O	O	svømmehal	O	O
fest	O	O	kirke	O	O	udenfor	O	O
forretning	O	O	park	O	O	zoologisk have	O	O
have	O	O	skole	O	O		O	O

12. Mennesker (30)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
baby	O	O	far	O	O	morfar	O	O
babysitters navn	O	O	farbror	O	O	mormor	O	O
barn	O	O	farfar	O	O	moster	O	O
barnets eget navn	O	O	farmor	O	O	oldefar	O	O
bedstefar	O	O	faster	O	O	olde(mor)	O	O
bedste(mor)	O	O	lærer	O	O	onkel	O	O
bror	O	O	mand	O	O	person	O	O
dagplejemors navn	O	O	menneske	O	O	pige	O	O
dame	O	O	mor	O	O	søster	O	O

dreng	O	O	morbror	O	O	tante	O	O

13. Leg og rutiner (15)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
aftensmad	O	O	hej hej (farvel)	O	O	shh	O	O
bade	O	O	ja	O	O	sove til middag	O	O
borte tit/titte bøh	O	O	klappe kage	O	O	tak	O	O
frokost	O	O	morgenmad	O	O	vent	O	O
hej	O	O	nej	O	O	vil	O	O

14. Ord om handlinger (52)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
bide	O	O	køre	O	O	slå	O	O
danse	O	O	lukke	O	O	smide	O	O
drikke	O	O	læse	O	O	sove	O	O
elske	O	O	løbe	O	O	sparke	O	O
falde	O	O	made	O	O	spise	O	O
få	O	O	plaske	O	O	standse	O	O
give	O	O	puste	O	O	svømme	O	O
græde	O	O	putte	O	O	synge	O	O
gynge	O	O	ride	O	O	sætte	O	O
gå	O	O	rydde op	O	O	tage	O	O
hjælpe	O	O	røre	O	O	tegne	O	O
holde øje med	O	O	se	O	O	trække	O	O
hoppe	O	O	sige	O	O	tørre af	O	O
kigge	O	O	skrive	O	O	vaske	O	O
kilde	O	O	skubbe	O	O	vise	O	O
kramme	O	O	skynde sig	O	O	ødelægge	O	O
kysse	O	O	[illegible]tte	O	O	åbne	O	O
lege	O	O	smile	O	O			

15. Ord, der beskriver (36)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
bange	O	O	i stykker	O	O	sulten	O	O
beskidt	O	O	kold	O	O	syg	O	O
blød	O	O	køn	O	O	sød	O	O
blå	O	O	lille	O	O	søvnig	O	O
fin	O	O	mørk	O	O	tom	O	O
forsigtig	O	O	(gør) ondt	O	O	træt	O	O
fræk	O	O	pæn	O	O	tør	O	O
gammel	O	O	rar	O	O	tørstig	O	O
glad	O	O	ren	O	O	ulækker	O	O
god	O	O	rød	O	O	varm	O	O
hurtig	O	O	slem	O	O	væk	O	O
hård	O	O	stor	O	O	våd	O	O

16. Ord om tid (8)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
dag	O	O	i morgen	O	O	nu	O	O
i aften	O	O	morgen	O	O	senere	O	O
i dag	O	O	nat	O	O			

17. Ord, der henviser (stedord) (11)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
den	O	O	din/dit/dine	O	O	jeg	O	O
denne/den her	O	O	du	O	O	mig	O	O
det	O	O	hans	O	O	min/mit/mine	O	O
dette/det her	O	O	hendes	O	O			

18. Spørgeord (6)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
hvad	O	O	hvor	O	O	hvorfor	O	O
hvem	O	O	hvordan	O	O	hvornår	O	O

19. Forholdsord og lokaliteter (16)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
af	○	○	inden i	○	○	ud	○	○
bagved	○	○	ned	○	○	ude	○	○
der	○	○	nede	○	○	under	○	○
i	○	○	op	○	○	væk	○	○
ind	○	○	oppe	○	○			
inde	○	○	på	○	○			

20. Kvantitetsord (kendeord m.v.) (10)

	forstår	forstår og siger		forstår	forstår og siger		forstår	forstår og siger
alle	○	○	ingen	○	○	noget	○	○
(en) anden	○	○	intet	○	○	samme	○	○
(et) andet	○	○	mere	○	○			
ikke	○	○	nogen/nogle	○	○			

Del II – Aktiviteter og gestik

A. FØRSTE KOMMUNIKATIVE GESTIKULATION

Når børn begynder at kommunikere, anvender de ofte bevægelser (gestik) for at gøre opmærksom på deres behov. For hvert spørgsmål nedenfor bedes du venligst markere det felt, der beskriver dit barns handlinger.

	ikke endnu	nogle gange	ofte
1. Strækker armen ud for at vise dig noget, som han/hun har i hånden	○	○	○
2. Rækker ud og giver dig et stykke legetøj eller anden genstand, som han/hun holder i hånden	○	○	○
3. Peger (med arm eller med pegefinger) på en interessant ting eller begivenhed	○	○	○
4. Vinker af sig selv farvel, når nogen tager afsted	○	○	○
5. Rækker sine arme op for at signalere, at han/hun gerne vil tages op	○	○	○
6. Ryster på hovedet for 'nej'	○	○	○
7. Nikker med hovedet for 'ja'	○	○	○

8. Markerer 'shh' ved at holde fingeren op foran munden	○	○	○
9. Beder om at få noget ved at række armen ud, mens han/hun åbner og lukker hånden	○	○	○
10. Laver fingerkys (på afstand)	○	○	○
11. Slikker sig om munden eller laver smaskelyde som tegn på, at noget smager godt	○	○	○
12. Slår ud med armene for at vise, at 'alt er væk' eller 'hvor blev det af'	○	○	○

B. LEG OG RUTINER

Gør dit barn noget af det følgende?

	ja	nej
1. Leger 'borte-tit-tit'	○	○
2. Leger 'klappe-kage'	○	○
3. Leger 'hvor stor er du?'	○	○
4. Leger 'nu skal jeg komme efter dig!'	○	○
5. Synger	○	○
6. Danser	○	○

C. HANDLINGER MED TING

Gør eller forsøger dit barn at gøre nogle af de følgende ting?

	ja	nej
1. Spiser med en ske eller gaffel	○	○
2. Drikker af en kop med væske i	○	○
3. Reder eller børster sit hår	○	○
4. Børster tænder	○	○
5. Tørrer ansigtet eller hænderne af med et håndklæde eller en klud	○	○
6. Tager hue på	○	○
7. Tager sko eller strømper på	○	○
8. Tager halskæde, armbånd eller ur på	○	○
9. Sænker hovedet og klemmer øjnene sammen for at lade som om han/hun sover	○	○
10. Puster for at vise, at noget er varmt	○	○
11. Holder en flyvemaskine og lader som om den flyver	○	○

	ja	nej
12. Holder telefonrøret op til øret	○	○
13. Lugter til blomster	○	○
14. Skubber til legetøjsbil	○	○
15. Kaster en bold	○	○
16. Lader som om hun/han hælder væske fra én beholder over i en anden	○	○
17. Lader som om der er væske i kopper eller gryder og rører rundt i det med en ske	○	○

D. LADER SOM OM HAN/HUN ER EN FORÆLDER

Her er nogle af de ting, som små børn sommetider gør med bamser og dukker. Markér venligst de handlinger, du har set dit barn udføre.

	ja	nej
1. Putter den i seng	○	○
2. Dækker den med et tæppe	○	○
3. Giver den sutteflaske	○	○
4. Mader den med en ske	○	○
5. Reder eller børster dens hår	○	○
6. Aer eller bøvser den	○	○
7. Skubber den i klapvogn eller dukkevogn	○	○
8. Vugger den	○	○
9. Kysser eller krammer den	○	○
10. Forsøger at give den sko, strømper eller hue på	○	○
11. Tørrer dens ansigt eller hænder	○	○
12. Snakker til den	○	○
13. Forsøger at give den ble på	○	○

E. IMITERER ANDRE VOKSENHANDLINGER (ved at bruge legetøj eller 'virkelige' ting)

Gør eller forsøger dit barn at gøre nogle af de følgende ting?

	ja	nej
1. Fejer med kost	○	○
2. Putter nøgle ind i en lås	○	○
3. Slår med en hammer eller lignende	○	○
4. Forsøger at bruge en sav	○	○
5. 'Skriver' på en skrivemaskine eller et computer-keyboard	○	○

6. 'Læser' (åbner bøger og vender sider)	○	○
7. Støvsuger	○	○
8. Vander blomster	○	○
9. Spiller på musikinstrumenter (fx klaver, trompet)	○	○
10. 'Kører' bil ved at dreje på et rat	○	○
11. Vasker op	○	○
12. Gør rent med en klud eller en støvkost	○	○
13. Skriver med en blyant, kuglepen, farve eller tusch	○	○
14. Graver med en skovl	○	○
15. Tager briller på	○	○

F. LADE SOM OM TING

Når børn leger, bruger de nogle gange ting som erstatning for andre. Et barn som gerne vil give sin bamse mad, lader måske som om en klods er et æble. Et barn kan også lade som om en skål er en hat. Har du set dit barn bruge erstatningsting på denne måde?

Hvis ja, giv eksempler:

ANDRE KOMMENTARER:

Appendix B. Sample scan of the Danish CDI Toddler part form *Ord og sætninger*

On the following pages, a sample scan of the Danish CDI Infant part form *Ord og gestikulation* is provided for illustration purposes only (version from February 29th, 2000). Under no circumstances it is permitted to use this sample as a basis for study conduct or other research purposes. See Bleses et al. (*in press*) for purchasing of the Danish CDI forms (University Press of Southern Denmark, Campusvej 55, 5230 Odense M, Tlf. +45 66 15 79 99, www.universitypress.dk).

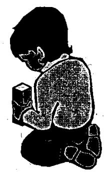
”We thank Larry Fenson and colleagues for permission to adapt the MacArthur-Bates Communicative Development Inventory to Danish. The adaptation into the Danish CDI was part of a larger project on language acquisition, *The Odense Language Acquisition-project* (1998-2001), funded by The Danish Research Council for the Humanities and University of Southern Denmark.” (Bleses et al., *submitted 2*, Acknowledgements)

For references, see chapter 1.

Barnets navn:__________________ **Fødselsdag:**________ **Dags dato:**_________.

MacArthur CDI.:

Ord og sætninger



The MacArthur Communicative Development Inventory:
Words and Sentences

For information/copies, contact The Developmental Psychology Lab, San Diego State University, San Diego, CA 92182

Tilpasset til dansk af

Odense-Projektet i Sprogtilegnelse,
Institut for Sprog og Kommunikation,
Syddansk Universitet, Odense Universitet 1999
Kontaktadresse: Campusvej 15, 5230 Odense M

Del I – Ord som børn anvender

A. ORDFORRÅD - CHECKLISTE

Børn forstår mange flere ord, end de siger. Vi er primært interesseret i de ord, som dit barn SIGER. Gennemgå venligst følgende ordliste og markér de ord, som du har hørt dit barn anvende. Hvis dit barn har en anden udtale af et ord (fx 'gugge' for 'dukke', 'dol' for 'stol' eller 'detti' for 'spaghetti'), bedes du alligevel markere ordet. Husk at dette er et 'katalog over ord', der kan findes hos mange forskellige børn. Vær ikke bekymret, hvis dit barn kun kender nogle få ord lige nu.

1. Lydeffekter og dyrelyde (12)

av	O	mjav	O	rap	O
fut	O	muh	O	vov	O
grrrr	O	mæh	O	åh åhh (ups)	O
kykliky	O	mm mm (lækkert)	O	årnnn (bil-lyd)	O

2. Dyrenavne (virkelige eller legetøj) (43)

abe	O	gås	O	løve	O
and	O	hane	O	mus	O
bamse	O	hest	O	myg	O
bi	O	hjort	O	myre	O
bjørn	O	hvalp	O	pingvin	O
dyr	O	hund	O	pony	O
egern	O	høne	O	skildpadde	O
elefant	O	kalkun	O	sommerfugl	O
fisk	O	kanin	O	tiger	O
flue	O	kat	O	ugle	O
frø	O	killing	O	ulv	O

fugl	O	ko	O	zebra	O
får	O	krokodille	O	æsel	O
giraf	O	kylling	O		
gris	O	lam	O		

3. Transportmidler (virkelige eller legetøj) (14)					
barnevogn	O	flyvemaskine	O	motorcykel	O
bil	O	helikopter	O	skib	O
brandbil	O	klapvogn	O	tog	O
bus	O	kælk	O	traktor	O
cykel	O	lastbil	O		

4. Legetøj (18)					
ballon	O	farveblyant	O	legetøj	O
blyant	O	gave	O	lego	O
bog	O	historie	O	puslespil	O
bold	O	klods	O	spil	O
dukke	O	kridt	O	sæbebobler	O
duplo	O	kuglepen	O	modellérvoks	O

5. Mad og drikke (68)					
agurk	O	kiks	O	sandwich	O
appelsin	O	kinderæg	O	slik	O
banan	O	kylling	O	slikkepind	O
boller	O	kød	O	smør	O
burger	O	leverpostej	O	småkager	O
brød	O	mad	O	sodavand	O

bønner	○	melon	○	sovs	○
chips	○	mælk	○	spaghetti	○
chokolade	○	mælkesnitte	○	spegepølse	○
cola	○	nødder	○	syltetøj	○
cornflakes	○	ost	○	suppe	○
dessert	○	ostehaps	○	te	○
fisk	○	pandekager	○	tun	○
frikadeller	○	pasta	○	tyggegummi	○
gulerødder	○	peanutbutter	○	vand	○
grød	○	pizza	○	vindruer	○
havregryn	○	pomfritter	○	vingummi	○
is	○	popcorn	○	vitaminer	○
jordbær	○	pølser	○	yoghurt	○
juice	○	rosiner	○	æble	○
kaffe	○	rugbrød	○	æg	○
kage(r)	○	saft(evand)	○	ærter	○
kartofler	○	salt	○		

6. Tøj (30)

ble	○	hue	○	strømpebukser	○
bluse	○	jakke	○	strømper	○
bukser	○	kjole	○	støvler	○
bælte	○	knap	○	sut	○
flyverdragt	○	kondisko	○	sweater	○
frakke	○	lynlås	○	t-shirt	○
hagesmæk	○	nattøj	○	trusser	○
halskæde	○	shorts	○	trøje	○
halstørklæde	○	sko	○	underbukser	○
hjemmesko/sutsko	○	sokker	○	vanter	○

7. Legemsdele (28)

ankel	○	hår	○	sår	○
ansigt	○	kind	○	tænder	○
arm	○	knæ	○	tissekone	○
ben	○	læber	○	tissemand	○
bryster	○	mave	○	tunge	○
finger	○	mund	○	tå	○
fødder	○	navle	○	øjne	○
hage	○	numse	○	ører	○
hoved	○	næse	○		
hånd	○	skulder	○		

8. Små husholdningsting (50)

affald	○	klokke	○	saks	○
bakke	○	kop	○	serviet	○
billede	○	kost	○	ske	○
briller	○	kniv	○	skrald	○
børste	○	krukke	○	skål	○
bånd	○	krus	○	småpenge	○
dyne	○	køkkenrulle	○	spand	○
dåse	○	lampe	○	støvsuger	○
flaske	○	lys	○	sæbe	○
gaffel	○	medicin	○	tallerken	○
glas	○	nøgler	○	tandbørste	○
gåvogn	○	papir	○	telefon	○
hammer	○	penge	○	toilet-/wc-papir	○
håndklæde	○	plante	○	tæppe	○
kam	○	pude	○	ur	○
kamera	○	pung	○	æske	○
kasse	○	radio	○		

9. Møbler og rum (33)					
altan	○	kælder	○	stue	○
badekar	○	køkken	○	terrasse	○
badeværelse	○	køleskab	○	toilet/wc	○
bord	○	ovn	○	trapper	○
bruser	○	potte	○	tv/fjernsyn	○
dør	○	seng	○	tørretumbler	○
garage	○	skab	○	vask	○
gyngestol	○	skuffe	○	vaskemaskine	○
høj stol	○	sofa	○	vindue	○
komfur	○	soveværelse	○	vugge	○
kravlegård	○	stol	○	værelse	○

10. Udendørsting (31)					
baggård	○	måne	○	sol	○
blomst	○	pind	○	sten	○
flag	○	græsslåmaskine	○	stige	○
fortov	○	regn	○	stjerne	○
gade	○	rutschebane	○	badebassin	○
græs	○	sandkasse	○	tag	○
gynge	○	skovl	○	træ	○
haveslange	○	ski	○	vand	○
himmel	○	slæde	○	vind	○
kælk	○	sne	○		
køkkenhave	○	snemand	○		

11. Steder (man kan tage hen til) (22)					
arbejde	○	fest	○	skole	○
(benzin)tanken	○	forretning	○	skoven	○
biografen	○	have	○	stranden	○

bondegård	○	hjem	○	svømmehal	○
butik	○	hus	○	udenfor	○
byen	○	kirke	○	zoologisk have	○
camping	○	legeplads	○		
cirkus	○	park	○		

12. Mennesker (40)

baby	○	farfar	○	moster	○
babysitters navn	○	farmor	○	oldefar	○
barn	○	faster	○	olde(mor)	○
barnets eget navn	○	folk	○	onkel	○
bedstefar	○	klovn	○	person	○
bedste(mor)	○	kæledyrets navn	○	pige	○
brandmand	○	læge/doktor	○	politi	○
bror	○	lærer	○	postbud	○
cowboy	○	mand	○	søster	○
dagplejemors navn	○	menneske	○	tante	○
dame	○	mor	○	ven	○
dreng	○	morbror	○	veninde	○
far	○	morfar	○		
farbror	○	mormor	○		

13. Leg og rutiner (27)

aftensmad	○	godnat	○	shh	○
bade	○	hej	○	sidde på potte	○
be om	○	hej hej (farvel)	○	"skal jeg komme efter dig"	○
borte tit/titte bøh	○	købe ind	○	snurre rundt	○
dav	○	ja	○	sove til middag	○
farvel	○	klappe kage	○	så stor!	○
frokost	○	morgenmad	○	tak	○
gider ikke	○	nej	○	vent	○
goddag	○	ringe	○	vil	○

14. Ord om handlinger (103)

arbejde	○	klatre	○	slikke	○
bide	○	kramme	○	slutte	○
blive	○	kysse	○	slå	○
bygge	○	købe	○	smage	○
bære	○	køre	○	smide	○
danse	○	lade som om	○	smile	○
dele	○	larme	○	snakke	○
drikke	○	lave mad	○	sove	○
elske	○	lege	○	sparke	○
falde	○	lide	○	spilde	○
fange	○	lytte	○	spille	○
feje	○	lukke	○	spise	○
finde	○	læse	○	stikke	○
få	○	løbe	○	standse	○
gemme	○	løbe efter	○	stå	○
give	○	made	○	svømme	○
glide	○	male	○	synge	○
græde	○	ordne	○	sætte	○
gynge	○	passe	○	tabe	○
gøre færdig	○	plaske	○	tage	○
gøre rent	○	puste	○	tegne	○
gå	○	putte	○	trække	○
havde	○	ride	○	tænke	○
have	○	rive i stykker	○	tørre	○
hente	○	rutche	○	tørre af	○
hjælpe	○	rydde op	○	vaske	○
holde	○	ryste	○	vente	○
holde øje med	○	røre	○	vise	○
hoppe	○	se	○	vælter	○
hælde	○	sidde	○	vågne	○
høre	○	sige	○	ødelægge	○

kaste	O	skrive	O	ønske	O
kigge	O	skubbe	O	åbne	O
kilde	O	skynde sig	O		
klappe	O	skære	O		

15. Ord, der beskriver (63)

bange	O	hård	O	stakkels	O
bedre	O	i stykker	O	stille	O
beskidt	O	ked af det	O	stor	O
blæser	O	klistret	O	sulten	O
blød	O	kold	O	sur	O
blå	O	køn	O	syg	O
brun	O	lang	O	sød	O
fin	O	langsom	O	søvnig	O
forsigtig	O	lille	O	sidst	O
forskrækket	O	mørk	O	tom	O
fræk	O	ny	O	træt	O
fuld	O	(gør) ondt	O	tung	O
først	O	orange	O	tør	O
gammel	O	pæn	O	tørstig	O
glad	O	rar	O	ulækker	O
god	O	ren	O	varm	O
grøn	O	rød	O	venlig	O
gul	O	sidde fast	O	vred	O
hurtig	O	slem	O	væk	O
hvid	O	snavset	O	våd	O
høj	O	sort	O	vågen	O

16. Ord om tid (15)

aften	O	før	O	morgen	O
dag	O	i aften	O	nat	O

efter	○	i dag	○	nu	○
eftermiddag	○	i går	○	senere	○
formiddag	○	i morgen	○	tid	○

17. Ord, der henviser (stedord) (31)

de	○	du	○	mig	○
dem	○	han	○	mig selv	○
den	○	hans	○	min/mit/mine	○
denne/den her	○	ham	○	os	○
dens/dets	○	hende	○	sig	○
deres	○	hendes	○	sig selv	○
det	○	hun	○	sin/sit/sine	○
dette/det her	○	I	○	vi	○
dig	○	jeg	○	vores	○
dig selv	○	jer	○		
din/dit/dine	○	jeres	○		

18. Spørgeord (7)

hvad	○	hvor	○	hvornår	○
hvem	○	hvordan	○		
hvilken/hvilket	○	hvorfor	○		

19. Forholdsord og lokaliteter (41)

af	○	hjemme	○	oppe	○
bagved	○	hos	○	over	○
bort	○	i	○	ovre	○
borte	○	ind	○	på	○
der	○	inde	○	til	○
for	○	inden i	○	tilbage	○
fra	○	i stedet for	○	ud	○

frem	○	med	○	ude	○
fremme	○	mod	○	uden for	○
gennem	○	ned	○	under	○
hen	○	nede	○	ved	○
henne	○	om	○	ved siden af	○
her	○	omme	○	væk	○
hjem	○	op	○		

20. Kvantitetsord (kendeord m.v.) (21)					
alle	○	det	○	mange	○
(en) anden	○	hver	○	meget	○
(et) andet	○	hvert	○	mere	○
andre	○	ikke	○	nogen/nogle	○
en (ikke tallet 1)	○	ingen	○	noget	○
et (ikke tallet 1)	○	intet	○	også	○
den	○	lidt	○	samme	○

21. Hjælpeudsagnsord (og mådesudsagnsord) (21)					
er	○	kan	○	skulle (datid)	○
gjorde	○	kunne (datid)	○	(at) skulle	○
gør	○	(at) kunne	○	var	○
gøre	○	må	○	vil	○
har	○	måtte (datid)	○	ville (datid)	○
ha(ve)	○	(at) måtte	○	(at) ville	○
havde	○	skal	○	være	○

22. Forbinderord (7)					
da	○	men	○	sådan	○
fordi	○	og	○		
hvis	○	så	○		

B. HVORDAN BØRN BRUGER ORD

Gør dit barn noget af det følgende?	ja	nej
1. Taler dit barn nogensinde om tidligere episoder eller personer, der ikke er tilstede? Fx kan et barn, der i sidste uge var i cirkus, senere sige "klovn", "elefant" eller "abe"	○	○
2. Taler dit barn nogen sinde om noget, der vil ske i fremtiden? Fx siger "fut-fut" eller "flyvemaskine", før I tager på en rejse eller siger "gyng-gang", før I går ud på legepladsen	○	○
3. Taler dit barn nogen sinde om ting, der ikke er tilstede? Fx spørger efter noget legetøj, som barnet ikke kan se, omtaler et kæledyr, der er uden for synsfeltet eller spørger efter en person, der ikke er til stede	○	○
4. Forstår dit barn, hvis du spørger efter noget, der ikke er tilstede? Fx ved at gå ind i stuen efter bamsen, mens du spørger: hvor er bamsen?	○	○
5. Peger dit barn på en genstand eller samler genstanden op og siger navnet på den person, som genstanden tilhører, også selvom personen ikke er tilstede? Fx ved at pege på mors sko og sige "mor"	○	○

Del II – Sætninger og grammatik

A. ORDS ENDELSER / DEL 1

	ikke endnu	nogle gange	ofte
1. Når vi taler om mere end en ting, føjer vi "-er" eller "-e" til mange ord, fx biler (for mere end én bil), bams**er**, dukk**er**, skib**e**, bold**e**. Er dit barn begyndt at gøre det?	○	○	○
2. Når vi taler om ejerforhold, tilføjer vi "-s", fx Peters bamse, bamsens ører, fars nøgler, mors bog. Er dit barn begyndt at gøre det?	○	○	○
3. Når vi taler om ting, der er sket i fortiden, tilføjer vi ofte "-ede" eller "-te" til udsagnsordet, fx snakk**ede**, hopp**ede**, spis**te**, kør**te**. Er dit barn begyndt at gøre det?	○	○	○

B. ORDFORMER

Her er nogle andre ord, børn lærer. Markér de ord, som dit barn bruger.

Navneord (flertalsformer)

børn	O	heste	O	mænd	O
fødder	O	hunde	O	skibe	O
(flere) får	O	(flere) mus	O	(flere) sko	O

Udsagnsord (datidsformer)

blev	O	gjorde	O	sad	O
blæste	O	havde	O	slog	O
drak	O	holdt	O	spiste	O
faldt	O	kom	O	så	O
fik	O	købte	O	tabte	O
fløj	O	kørte	O	tog	O
gik	O	løb	O		

C. ORDS ENDELSER / DEL 2

Børn tilføjer ofte de forkerte endelser til ord. Fx kan et barn finde på at sige "moster gåede hjem". Den slags "fejl" er tegn på en sproglig udvikling. Markér venligst de "fejl", der minder om dem, dit barn har sagt i den seneste tid.

Navneord (flertalsformer)

(flere) "barn"	O	"fodder"	O	"mænder"	O
"barne"	O	"fød"	O	"muse"	O
"barner"	O	"fødde"	O	"muser"	O
"børne"	O	(flere) "mand"	O	"skoe"	O
"børner"	O	"mande"	O	"skoer"	O
(flere) "fod"	O	"mander"	O	"tander"	O
"fodde"	O	"mænde"	O	"tåer"	O

Udsagnsord (datidsformer)

"bringede"	O	"gørte"	O	"sete"	O

”bringte”	○	”gåede”	○	”siddede”	○
”blæsede”	○	”gåte”	○	”sidte”	○
”drakkede”	○	”havede”	○	”såddede”	○
”drakte”	○	”havte”	○	”såte”	○
”drikkede”	○	”holdede”	○	”tagede”	○
”drikte”	○	”holdte”	○	”tagte”	○
”faldede”	○	”hørede”	○	”togede”	○
”faldte”	○	”kommede”	○	”togte”	○
”flyvede”	○	”komte”	○	”ødelaggede”	○
”flyvte”	○	”løbede”	○	”ødelagte”	○
”fåede”	○	”løbte”	○	”ødelæggede”	○
”fåte”	○	”sadte”	○	”ødelægte”	○
”gørede”	○	”seede”	○		

Er dit barn begyndt at kombinere ord, som fx “mere kiks” eller “hund bider”?

ikke endnu	indimellem	ofte
○	○	○

Hvis du har afkrydset ‘*ikke endnu*’-bokse, så stop her. Hvis du har sat kryds i én af de andre bokse, så besvar venligst de næste spørgsmål.

D. EKSEMPLER

Skriv venligst tre af de længste sætninger, du har hørt dit barn sige fornylig.

1. ______________________________

2. ______________________________

3. ______________________________

E. KOMPLEKSITET

Markér venligst den af følgende sætningspar, der lyder mest som det dit barn siger.

1. To bil To biler	12. Det min bil Det er min bil	23. Tænd lys Tænd lyset så jeg kan se
2. To fod To fødder	13. Du ordne det? Kan du ordne det?	24. Jeg vil ha' den Jeg vil ha' den som du har
3. Far bil Fars bil	14. Læs historie mor Læs en historie for mig mor	25. Vil have kiks Vil have kiks og mælk
Ved tale om noget, der allerede er sket 4. Jeg falder ned Jeg faldt ned	15. Nej vaske dukken Ikke vaske dukken	26. Kiks mor Kiks til mor
5. Flere småkage Flere småkager	16. Ha' mere saft Ha' (mere) saft i den her	27. Baby vil spise Baby vil gerne spise
6. De (er) min tand De (er) mine tænder	17. Der en mis Der er en mis	28. Se mig Se mig danse
7. Baby tæppe Babys tæppe	18. Hvor mor gå hen? Hvor gik mor hen?	29. Se Se hvad jeg har
Ved tale om noget, der allerede er sket 8. Hunden kysser mig Hunden kyssede mig	19. Kaffe varm Den kaffe varm	30. Hvor er min dukke? Hvor er min dukke, der hedder Lise?
Ved tale om noget, der allerede er sket 9. Far tager mig op Far tog mig op	20. Jeg nej gør det Jeg kan ikke gøre det	31. Vi lavede det Mig og Paul lavede det
Ved tale om noget, der allerede er sket 10. Katten går væk Katten gik væk	21. Jeg kan lide læse historier Jeg kan lide at læse historier	32. Jeg synger en sang Jeg synger en sang for dig
11. Hunden bord Hunden på bordet	22. Ikke læse bog Jeg vil ikke (ha) du læse den bog	33. Baby græder Baby græder fordi hun er ked af det

ANDRE KOMMENTARER: