J. Child Lang. **37** (2010), 419–428. © Cambridge University Press 2009
doi:10.1017/S0305000909009544

BRIEF RESEARCH REPORT

CLEX: A cross-linguistic lexical norms database*

RUNE NØRGAARD JØRGENSEN
University of Southern Denmark

PHILIP S. DALE
University of New Mexico

DORTHE BLESES
University of Southern Denmark

AND

LARRY FENSON
San Diego State University

(*Received* 11 *October* 2008 – *Revised* 2 *February* 2009 – *Accepted* 18 *February* 2009 – *First published online* 2 *July* 2009)

ABSTRACT

Parent report has proven a valid and cost-effective means of evaluating early child language. Norming datasets for these instruments, which provide the basis for standardized comparisons of individual children to a population, can also be used to derive norms for the acquisition of individual words in production and comprehension and also early gestures and symbolic actions. These lexical norms have a wide range of uses in basic research, assessment and intervention. In addition, cross-linguistic comparisons of lexical development are greatly facilitated by the availability of norms from diverse languages. This report describes the development of CLEX, a new web-based cross-linguistic database for lexical data from adaptations of the MacArthur-Bates Communicative Development Inventories. CLEX provides tools for a range of analyses within and across languages. It is designed to incorporate additional language datasets easily, and to permit users to define

[*] Address for correspondence: Rune Nørgaard Jørgensen, Center for Child Language, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. e-mail: rune@sdu.dk

mappings between lexical items in pairs of languages for more specific cross-linguistic comparisons.

Parent report has proven a valid and cost-effective means of evaluating early child language (Dale, 1996; Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007). Its validity rests on the far larger and more representative experience parents can have with their child's language than even an expert can obtain in standardized testing or language sampling. Furthermore, the ability to obtain large datasets relatively inexpensively has made possible the construction of norms for early language development that are significantly better than were previously available, thus adding substantially to the precision of our evaluation of individual children. For the same reasons, parent report is highly useful for research. It is particularly well-suited for research that requires large samples, such as twin studies on the influence of genetic and environmental variables on language development (e.g. Plomin & Dale, 2000), and research on the effects of specific environmental factors such as daycare and television on language development (e.g. NICHD Early Child Care Research Network, 2000).

The most widely used parent report measures of early language development in English are the MacArthur-Bates Communicative Development Inventories, often referred to as CDIs (Fenson *et al.*, 2007). In addition to the assessment of typically developing children, they have been shown to provide valid information on atypical populations for child language development, such as Down syndrome, Specific Language Impairment, hearing impairment and autism (Fenson *et al.*, 2007). Because the CDIs have proved so useful, they have been, or are in the process of being, adapted into more than forty-five languages (see the CDI website, www.sci.sdsu.edu/cdi/ for a listing of projects underway or completed). Both linguistic and cultural adaptations must be made for gestures, words and grammatical structures in the development of these new measures. However, as the core structure of the measures is relatively similar across languages, cross-linguistic comparisons can be made on the basis of larger and more representative samples than can be done for more labor-intensive methods such as those that require language sampling. For example, Bleses *et al.* (2008*a*) identified numerous common aspects of development in seventeen languages for which adapted CDI norming data were available. They also noted that, with respect to early lexical comprehension, Danish was notably slower than other languages, even than Swedish, a closely related language. They attributed this difference to some unique properties of Danish phonology which could be expected to make word segmentation difficult.

The norming datasets which provide the basis for standardized comparison of individual children to a population can also be used to derive norms for the acquisition of individual words in production and

comprehension. Instead of aggregating words for individual children, we can aggregate children for individual words. In this way, Fenson, Dale, Reznick, Bates, Thal & Pethick (1994) established age of acquisition norms for individual words on the CDI: Words & Gestures, separately for comprehension and production, and for individual words produced on the CDI: Words & Sentences, based on the age at which at least 50 percent of parents provide a positive answer to the relevant question. A similar analysis can be done for gestures and symbolic actions. Evidence for the validity and utility of these lexical norms comes from the findings of Goodman, Dale & Li (2008), who showed a significant relation between frequency of individual words in parental input (based on CHILDES transcripts) and age of acquisition derived from the CDI, when the correlations were conducted WITHIN specific form classes such as nouns, verbs and closed class words.

In 1995, the CDI Advisory Board developed a database with month-by-month norms for individual lexical items in the norming dataset for the CDI. This was made available first as the standalone program LEX (Dale & Fenson, 1996) and later as a web-based application on the CDI website. The present paper reports on the development of an extension of LEX, entitled CLEX, designed specifically for cross-linguistic research and application. CLEX is an acronym for Cross-linguistic Lexical Norms. It has been developed collaboratively by the CDI Advisory Board and the Center for Child Language, University of Southern Denmark, and is presently hosted by the University of Southern Denmark at www.cdi-clex.org. In this paper we describe some example applications of this lexical norm information, give a brief overview of CLEX functionality, and provide an overview on adding new information – additional language datasets and cross-linguistic mappings – to CLEX.

APPLICATIONS OF LEXICAL NORMS

Developmental normative information on individual lexical items and sets of items has multiple uses in research and clinical application. Information on individual lexical items provides an empirical basis for selecting words which can be assumed to be very likely known at a given age, and therefore appropriate, for example, for use in sentences testing grammatical or pragmatic development. This information may be equally valuable in research in other areas, such as cognitive or social development, when there is a need to equate words and even pictures or objects for estimated familiarity. Conversely, words can be selected as very likely not to be known at a given age; these words can be used in studies of lexical learning, or of neurophysiological response to known vs. unknown words. Developmental norms can also provide a basis for selection of targets for clinical intervention on the basis of words which are likely to be learned next in

typical development, relative to a child's current status. The nature of growth of individual words in the population is also illuminating. For example, Fenson *et al.* (1994) classified words into three categories with respect to the linear, quadratic and cubic components of their growth curve, and offered a psycholinguistic interpretation of the results on the basis of holistic (or formulaic) processes of development being more likely for some categories of words than others.

A second level at which norms can be used is the comparison of specific pairs and larger sets of words for evaluating theory-driven hypotheses. For example, are positive members of antonym pairs of dimensional terms consistently acquired earlier than the negative member, e.g. 'big' before 'little'? And what is the order of acquisition of kinship terms?

In these first two levels of application, developmental information on individual words is the basic unit of analysis. The CDI instruments have lexical items organized into categories (nineteen for the CDI:WG and twenty-two for the CDI:WS, but this differs across languages) which are largely semantic in nature, though they usually have some syntactic coherence as well, e.g. toys, food and drink, actions, time words, etc. Total scores for these categories are widely used in studies in vocabulary composition, along with still broader aggregations, such as common nominals, predicates and closed class words (Bates *et al.*, 1994). However, an individual researcher may have a rationale for a different category, for example, actions which are punctate vs. extended in time, words with an overall positive tone, etc. The development of that category, perhaps in contrast to a related category, may be the focus of interest. Similarly, it is often useful in developmental research to match children on overall vocabulary development prior to an experimental intervention or assessment. It may be even more useful to do the matching on the vocabulary category of special interest in that study.

In most analyses, the developmental information on individual items is based on means, that is, the average age of acquisition of words, or the percentage of children who have produced the word by, e.g. age 2;0. Furthermore, the norming data have most often been collected in cross-sectional studies. Consequently, the types of analysis just mentioned do not make it possible to say that word A is always learned before word B, only that it is learned first on average. However, some kinds of longitudinal hypotheses can be evaluated with cross-sectional data using scaling analyses, if access to individual datapoints – each word, for each child in the norming sample – is available. For example, if it is hypothesized that word A is always learned before word B, we can predict that within a cross-sectional sample children can be identified who have mastered both words, neither word, or word A but not word B; the pattern of word B without word A should not be seen.

As mentioned earlier, adaptations of the CDI have made possible cross-linguistic comparisons based on large and representative samples. All three types of research just mentioned, examining individual words known or not known, user-defined subscales and the relationships among words at the level of individual children, are useful in cross-linguistic research. When lexical items can be matched across languages, a comparison of their age of acquisition – both in absolute terms of age, and relative to other words – with differences in phonological, morphological or syntactic properties can be valuable for evaluating theoretical proposals concerning acquisition mechanisms. The same is true for categories of words; for example, Tardif, Fletcher, Liang & Kaciroti (in press) have shown that classifiers emerge later in age for Cantonese than Mandarin speakers, but earlier with respect to vocabulary size, reflecting grammatical differences between the two Chinese languages with respect to classifiers.

CLEX FUNCTIONALITY

All queries in CLEX begin with selection of the primary dataset; that is, the language, the form (CDI:WG or CDI:WS) and, for CDI:WG, whether it is comprehension or production that is of interest. Most datasets, but perhaps not all, will have the two forms, though that is not necessary for inclusion in CLEX. It is also possible that eventually there will be multiple datasets for the same instrument in a given language, e.g. one from a cross-sectional study and another from a longitudinal study. In addition, the age span may vary across languages, e.g. the Danish CDI:WG is normed from age 0;8–1;8, whereas the range for the US English norming data is age 0;8–1;6. It should also be noted that for almost all analyses, once in the relevant portion of the program it is possible to restrict the dataset still further, by gender or by age. In the longer term a selection parameter of total vocabulary size is planned as well.

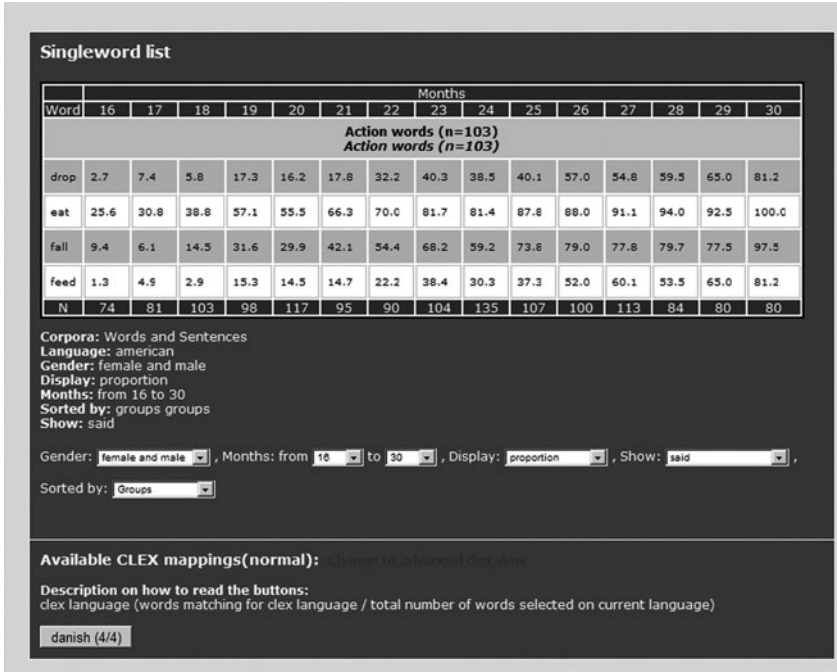
The 'main menu' then offers five types of analysis.

Norms. This option provides the overall normative information on vocabulary totals by age comparable to the tables and figures in the Technical Manuals for the instruments.

Single Word. The next four options are the core of CLEX. Under the 'Single Word' heading, the user selects an individual word and output is provided, in tabular or graphical form, of the developmental increase in the number of children who are reported to produce or understand the word.

Single Word List. In this option, the user can specify a list of words, and information about the development of each one is provided separately, but simultaneously, in a table. Figure 1 provides sample output for such an analysis.

https://doi.org/10.1017/S0305000909009544 Published online by Cambridge University Press



Singleword list

Word	Months 16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Action words (n=103) *Action words (n=103)*															
drop	2.7	7.4	5.8	17.3	16.2	17.8	32.2	40.3	38.5	40.1	57.0	54.8	59.5	65.0	81.2
eat	25.6	30.8	38.8	57.1	55.5	66.3	70.0	81.7	81.4	87.6	88.0	91.1	94.0	92.5	100.0
fall	9.4	6.1	14.5	31.6	29.9	42.1	54.4	68.2	59.2	73.8	79.0	77.8	79.7	77.5	97.5
feed	1.3	4.9	2.9	15.3	14.5	14.7	22.2	38.4	30.3	37.3	52.0	60.1	53.5	65.0	81.2
N	74	81	103	98	117	95	90	104	135	107	100	113	84	80	80

Corpora: Words and Sentences
Language: american
Gender: female and male
Display: proportion
Months: from 16 to 30
Sorted by: groups groups
Show: said

Gender: female and male , Months: from 16 to 30 , Display: proportion , Show: said ,

Sorted by: Groups

Available CLEX mappings(normal): Change to advanced clex view

Description on how to read the buttons:
clex language (words matching for clex language / total number of words selected on current language)

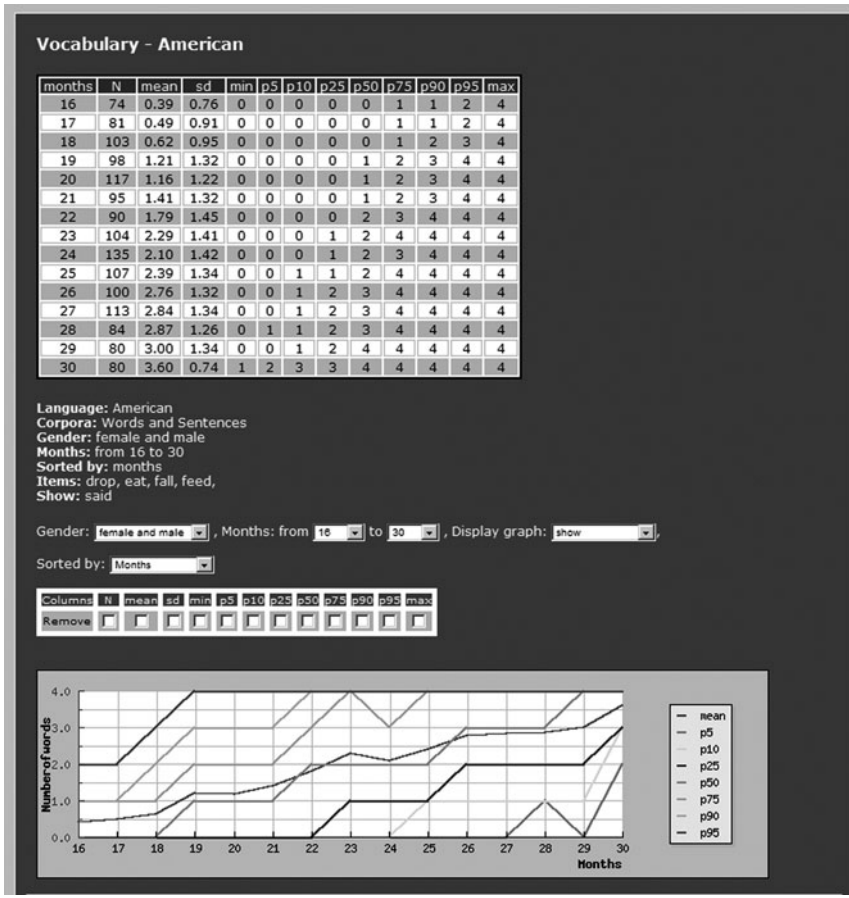
danish (4/4)

Fig. 1. CLEX output for analysis of Single Word List of four user-selected words.

Vocabulary Subscale. Here the user specifies a list of words, but rather than analyzing them separately, CLEX treats the list as a vocabulary subscale, and reports the percentiles, mean, min, max and SD over time. Figure 2 displays sample output; in this case, a graph was requested as well as tabled output.

Direct Item Comparison. All of the above analyses are based on data aggregated by age in months. For many questions, as discussed above, it is the direct comparison of items which is most relevant. For example, the Single Word List option would allow the determination that *fall* is generally learned before *drop*. But a direct comparison is needed to determine if, for those children who have just one of these two words, it is always, or nearly always, *fall*. In the Direct Item Comparison analysis, the user specifies a list of words (maximum = 6) and CLEX reports the frequency of all possible patterns of those words, as shown in Figure 3. That table may be exported to Excel or a statistical program for further analysis. We anticipate that further types of direct item analysis will be added to the system in the future.

https://doi.org/10.1017/S0305000909009544 Published online by Cambridge University Press



months	N	mean	sd	min	p5	p10	p25	p50	p75	p90	p95	max
16	74	0.39	0.76	0	0	0	0	0	1	1	2	4
17	81	0.49	0.91	0	0	0	0	0	1	1	2	4
18	103	0.62	0.95	0	0	0	0	0	1	2	3	4
19	98	1.21	1.32	0	0	0	0	1	2	3	4	4
20	117	1.16	1.22	0	0	0	0	1	2	3	4	4
21	95	1.41	1.32	0	0	0	0	1	2	3	4	4
22	90	1.79	1.45	0	0	0	0	2	3	4	4	4
23	104	2.29	1.41	0	0	0	1	2	4	4	4	4
24	135	2.10	1.42	0	0	0	1	2	3	4	4	4
25	107	2.39	1.34	0	0	1	1	2	4	4	4	4
26	100	2.76	1.32	0	0	1	2	3	4	4	4	4
27	113	2.84	1.34	0	0	1	2	3	4	4	4	4
28	84	2.87	1.26	0	1	1	2	3	4	4	4	4
29	80	3.00	1.34	0	0	1	2	4	4	4	4	4
30	80	3.60	0.74	1	2	3	3	4	4	4	4	4

Fig. 2. CLEX output for analysis of Vocabulary Subscale composed of the same four words.

Cross-linguistic Comparisons. The availability of comparable data from CDI adaptations across languages within CLEX makes possible a wide range of cross-linguistic research. The four main analysis options, Single Word, Single Word List, Vocabulary Subscale and Direct Item Comparison, all offer the possibility of proceeding to comparable analyses with the related items in another language, if they exist on the other form. For example, *slaede* ('sledge') is on the Danish CDI, but not on the American one, so no comparison is possible. Sample output for the words expressing the concept of 'mother' in English and Danish is shown in Figure 4.

Considerable caution is advised in taking advantage of the possibility of cross-linguistic comparisons. A default system of mappings between

https://doi.org/10.1017/S0305000909009544 Published online by Cambridge University Press

Fig. 3. CLEX output for Direct Item Comparison: a cross-tabulation of responses to two words.

Fig. 4. CLEX output for comparison of two words, one from English and one from Danish.
NOTE: A color version of Figures 1–4 can be found at http://journals.cambridge.org/jcl.

American English and each non-English language is provided, usually by the investigator who conducted the norming study and has contributed the data. That system of mappings is useful for general orientation within the language and identifying items of particular interest, particularly for

https://doi.org/10.1017/S0305000909009544 Published online by Cambridge University Press

non-speakers of the language. But because exact synonymy is seldom found across languages, users must consider their research questions and hypotheses in determining what is to count as a mapping. Cognate status, phonological similarity, degree of semantic relatedness and other factors may be very important for specific projects. For this reason, CLEX provides a functionality by which either the initial developer/contributor of the data or a later researcher can input a distinct system of mappings for their own or others' use. At present, American English serves as the 'interlanguage' and comparisons between other languages are made via their link to it. In the future, direct mappings between non-English languages may be added to the system. We also anticipate adding the capacity for datasets for bilingual children, which would permit the investigation of hypotheses concerning both the rate and nature of bilingual lexical development.

A *CLEX User's Guide* in pdf format has been prepared and posted on the website. We invite comments and suggestions for improving its usefulness. The website also includes links to several documents with information on obtaining authorization for developing new CDI adaptations, and suggestions for the process.

'GROWING' CLEX

The value of CLEX will grow as the number of included languages increases. At present, the system includes data from American English, Danish and Swedish. More languages are about to follow. We invite other investigators who have developed an adaptation of the CDI to join in this project. All data are to be transferred anonymously; only age and gender are required information about the participants. The CLEX development team (info@cdi-clex.org) will work with investigators to prepare datasets, and obtain other necessary information such as names of semantic subcategories and the set of initial mappings of individual items to American English. We also anticipate adding datasets from specific atypical populations, such as Specific Language Impairment and Down Syndrome, when they are of adequate size and representativeness.

A major inspiration for the development of CLEX has been the success of the Child Language Data Exchange System (CHILDES; MacWhinney, 2000) in promoting child language research. One crucial factor in the success of CHILDES has been the generosity of researchers in contributing data, and we hope this will extend to CDI data for CLEX. Like CHILDES, we are developing, and posting on the website, explicit policies concerning acknowledgment of the use of data. Another contributing factor has been the ingenuity and insights of the CHILDES development team and many others in the scientific discipline. We hope that researchers will

contribute to CLEX not only CDI datasets, but also suggestions for analytic tools which can facilitate research.

SUPPLEMENTARY MATERIAL: the supplementary material in the article can be found at http://journals.cambridge.org/jcl.

REFERENCES

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**, 85–123.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008*a*). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language* **35**, 619–50.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008*b*). The Danish Communicative Development Inventories: validity and main developmental trends. *Journal of Child Language* **35**, 651–69.

Dale, P. S. (1996). Parent report assessment of language and communication. In K. N. Cole, P. S. Dale & D. J. Thal (eds), *Assessment of communication and language*, 161–82. Baltimore: Paul H. Brookes Publishing Co.

Dale, P. S. & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers* **28**, 125–27.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J. & Pethick, S. J. (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. J., Dale, P. S., Reznick, J. S. & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories. Users guide and technical manual* (2nd edn). Baltimore: Paul H. Brookes Publishing Co.

Goodman, J. C., Dale, P. S. & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language* **35**, 515–31.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

NICHD Early Child Care Research Network (2000). The relation of child care to cognitive and language development. *Child Development* **71**, 960–80.

Plomin, R. & Dale, P. S. (2000). Genetics and early language development: A UK study of twins. In D. V. M. Bishop & L. B. Leonard (eds), *Speech and language impairments in children: Causes, characteristics, intervention, and outcome*, 35–51. Philadelphia: Taylor & Francis.

Tardif, T., Fletcher, P., Liang, W. & Kaciroti, N. (in press). Early vocabulary development in Mandarin (Putonghua) and Cantonese. *Journal of Child Language*.