Statistical Learning in Emerging Lexicons

The Case of Danish

Stokes, Stephanie; Bleses, Dorthe; Basbøll, Hans; Lambertsen, Claus

Published in:
Journal of Speech, Language, and Hearing Research

DOI:
10.1044/1092-4388(2012/10-0291)

Publication date:
2012

Document version:
Accepted manuscript

Citation for pulished version (APA):
Stokes, S., Bleses, D., Basbøll, H., & Lambertsen, C. (2012). Statistical Learning in Emerging Lexicons: The Case of Danish. *Journal of Speech, Language, and Hearing Research*, *55*, 1265-1273. https://doi.org/10.1044/1092-4388(2012/10-0291)

Statistical Learning in Emerging Lexicons: The Case of Danish

Stephanie F. Stokes[a], Dorthe Bleses[b], Hans Basbøll[b], and Claus Lambertsen[b]

[a]University of Canterbury, New Zealand, [b]Center for Child language, University of Southern Denmark

Contact: Stephanie Stokes (stephanie.stokes@canterbury.ac.nz), Department of Communication Disorders, University of Canterbury, Private Bag 4800, New Zealand, 8140.

Key words: statistical learning, late talkers, phonological neighborhood density, Danish

Abstract

Purpose: This research explored the impact of neighborbood density (ND), word frequency (WF), and word length (WL) on the lexicons of Danish-speaking children. Given the particular phonological properties of Danish the impact was expected to differ from reports on English and French.

Method: The monosyllabic words in the expressive lexicons of 894 Danish-speaking two-year-old children were coded for ND, WF and WL. Lexicons were extracted from parent checklists of the words spoken by their children.

Results: Regression revealed that ND, WF, WL and Age together predicted 47% of the variance in vocabulary size, with ND, WF, WL and Age uniquely accounting for 39%, 3.2%, 2.2% and 2.8% of that variance respectively. Between-group comparisons showed that children with small vocabularies had learned words that were denser, more frequent, and shorter than children with average or large vocabularies.

Conclusion: The two main findings were unexpected. The impact of ND for Danish-speaking children was not expected given the phonological properties of the language. Further, the morphological structure of verbs generated a surprising result for WF. The strong role for ND in emerging languages found in other languages was replicated for Danish.

This research explores the cross-linguistic validity of a recent claim made in the field of child language development. The claim is that children who are struggling to learn their ambient language are learning words that are comprised of statistical properties that differ from those used by their more able peers (e.g., Stokes, 2010; Stokes, Kern & dos Santos, 2011). Children who are struggling to learn a lexicon and who meet certain criteria are described as 'late talkers' (LTs). The criteria are having have had a slow onset of expressive vocabulary and slow vocabulary growth in the second year of life, resulting in small lexicons at two years relative to their age-matched TD peers, (e.g. Demarais, Sylvestre, Meyer, Bairati & Rouleau, 2008). The two most commonly used quantitative metrics are less than 50 words or no word combinations at 24-30 months (e.g., Paul, 1996), or being at or below the 10th percentile on the MacArthur-Bates Communicative Development Inventory (MCDI; Fenson, Dale, Reznick, Thal, Bates, Hartung, et al., 1993; Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007) relative to an age-matched cohort (e.g., Bishop, Dale & Plomin, 2003).

In addition to these quantitative metrics, qualitative differences in LTs vocabularies relative to their peers have recently been identified. These differences were in the lexical characteristics of words, neighborhood density (ND) and word frequency (WF) specifically. In studying the MCDI expressive vocabularies of 220 English- and 208 French-speaking two-year-old children Stokes and colleagues (Stokes, 2010; Stokes et al., 2011), reported that ND accounted for 47% and 53% of the variance in vocabulary size for the English-speaking and French-speaking children respectively. As vocabulary size increased, ND values decreased, reflecting the increase in learning words from sparse phonological neighborhoods. The findings for ND were congruent with prior reports, using norming data from the MCDI, that the earliest

learned words came from dense phonological neighborhoods in the ambient language when (e.g., Storkel, 2004a, 2008a, 2008b).

Surprisingly, word frequency values accounted for only a small but significant amount of variance in vocabulary size once ND had been accounted for (14% for English and 9% for French). The direction of the relationship was positive. Small vocabularies were comprised of words that were less frequent in the ambient input and large vocabularies were comprised of more frequent words. These WF results were at odds with many reports that earliest learned words were of high frequency in the input. However, they were in agreement with reports that the patterns for frequency did not hold across all children (Storkel, 2004a, 2008a), and that the patterns varied as a function of word class (Goodman, Dale & Li, 2008). Up until 2008, it was generally accepted that children first learned (spoke) high frequency words, however, Goodman et al's (2008) large-scale analysis of the relationship between word frequency in the input language and age-of-acquisition of words in children's productive lexicons showed that when all word classes were examined together, there was a strong positive relationship between age of acquisition in production and the frequency of words in CDS (words learnt earliest were of the lowest frequency in the input). The six lexical categories studied were common nouns, people words, verbs, adjectives, closed class and others. (Common nouns were words that encoded objects and substances, like car, dog and milk, and nouns that labeled locations or events, like beach and lunch were categorized as other.) When word classes were examined separately (e.g., only nouns), word frequency was negatively correlated with age of acquisition, indicating that words learnt earliest were of high frequency. The Goodman et al (2008) and Stokes et al (2010, 2011) results raised the possibility that the conventional wisdom of high frequency words being learnt first might not hold for all children or all word classes.

If ND is a strong cue in developing an expressive lexicon, then this cue should operate across languages. Note that the ND values for English words were derived from a combination of rhyme neighbors (e.g. cat, hat, mat), lead neighbors (e.g. hat, ham, have) and consonants neighbors (e.g. hot, hat, hut). As with other research on this topic (e.g., Storkel, 2004a; Zamuner, 2008), Stokes and colleagues summed all of these types into the category of phonological neighbor. However, as languages vary by the very characteristic that defines ND (the number of rhyme, lead and consonant neighbors as a function of the number of words comprised of $[C_i+V_1+C_1]$, $[C_1+V_1+C_i]$ and $[C_1+V_i+C_2]$ respectively) then not all languages may reflect the patterns identified for English and French. Languages differ with respect to the numbers of consonants and vowels (and the resulting consonant/vowel ratio), phonotactic structures, and the lengths of words, and so children learning languages that differ on these prime characteristics may not show the same learning patterns that were identified for English and French. If they do show the same patterns for ND and WF then some cross-linguistic learning heuristic (see below) can be hypothesised. One language that differs from English and French on the relevant phonological properties, and for which a database was available, was Danish.

Danish

French and English differ on some phonological properties but the differences are quite small (Table 1). There is little difference between the languages in the number of consonants and vowels, and little difference in the C/V ratio, with both around 1.5, indicating that consonant and vowel effects on lexical processing are likely to be similar across these two languages (Carreiras & Price, 2008). As mentioned above, the relationship between ND and vocabulary size was very similar for English and French. However, Danish is a language that has considerably more vowels than consonants (Table 1). Carreiras and Price (2008) claimed that consonants play more

of a role in lexical processing than vowels and therefore languages that have considerably fewer consonants than vowels should pose a different learnability problem than has been seen for English and French. Danish has 18 consonants and 37 vowels, yielding a C/V ratio of .49. If consonants play a more important role than vowels in the detection of word boundaries in a continuous speech stream (Bonatti, Peña, Nespor, & Mehler, 2004), then children learning Danish would seem to be at a disadvantage if ND is a primary statistical cue to lexical learning. English and French may provide more ND learning opportunities because the number of rhyme, lead and consonant neighbors in monosyllabic words must vary as a function of the relative number of consonants and vowels in a language.

Table 1 about here

The second factor important here is that of phonotactic properties. Descriptions of Danish report that a) word boundaries and syllables are indistinct and b) syllable reduction in Danish generates highly variable word forms for any given lexeme, although syllable reduction is more likely to occur in disyllabic than monosyllabic words (Basbøll, 2005; 2006). Basbøll used an example from Rischel (2003), of how the phrase '(h)årdere at åre(lade)' 'harder to bleed' (with reference to a zoo elephant being harder to bleed than others for the veterinarian) is realized as one overlong vowel segment [ɒ] distributed over six syllables /ɒ:ɒ ɒ ɒɒ:ɒ/. The process is not just one of phrasal coarticulation as happens in all languages. Consonant gradation in Danish (see Rischel, 1970) occurs in isolated words also, such that obstruents in word-final position or before a schwa phoneme are realized as non-lateral approximants. In connected speech, complete elision with syllable loss, consonant gradation and schwa reduction (where a schwa is assimilated to a neighboring sonorant consonant, yielding a syllabic sonorant) results in long vowel-like strings (cf. Basbøll, 2005). The consequences of this phonetic structure is that Danish

children have to find boundaries in vowel sequences which blur syllable boundaries word-internally as well as word-externally and that possibly makes not only word segmentation more difficult than would be the case for other languages but also the acquisition of inflectional morphology (Grønnum, 2003; Bleses, Vach, Slott, Wehberg, Thomsen, Madsen & Basbøll, 2008a; Bleses, Basbøll & Vach, 2010; Bleses, Basbøll, Lum & Vach, 2010). For children attempting to develop a lexicon, employing statistical learning strategies that result in learning words with high statistical probabilistic cues (words from dense neighborhoods in the ambient language) will only succeed where such cues are available. An alternative learning device would be required where ND cannot be employed.

Alternative probabilistic cues

One such alternative device may be word frequency (WF). While WF played a minor role in accounting for the variance in vocabulary size in English and French, this may not be the case for Danish where ND is possibly harder to employ as a learning cue. Word length is also known to impact on word learning, with shorter words being learned earlier than longer words, for both child data from norming datasets (e.g., Storkel, 2004; Maekawa & Storkel, 2006), and from network simulations of lexical learning (e.g., Li & Zhao, 2007). It is also clear that word length is significantly correlated with ND, such that short words have more neighbors than long words (Storkel, 2004b). It is tempting then to suggest that children (and simulations) simply learn short words, and that learning words from dense neighborhoods is an artifact. However, a question of logic arises, one that can be demonstrated by considering Li & Zhao's (2007) results. For both English and Chinese, the network simulation learned short words before learning long words. In both languages, there is a predominance of short words in the language. Chinese verbs are predominantly two to four phonemes in length, with 40% of verbs being comprised of three

phonemes, nouns are mostly three to six phonemes, with 25% comprised of four phonemes. In English, word length for nouns and verbs is similar, mostly between three and five phonemes. If short words predominate, any system would learn more short words than long words, regardless of factors such as ease of articulation or lower impact on short-term memory. This being so, it becomes important not to make an A-priori decision about whether or not other variables (such as ND) should be standardized as a function of word length. Rather, working without a fixed A-priori hypothesis, all variables would be entered equally (via forced entry) into a regression analysis to determine how much of the variance in vocabulary size can be attributed jointly and uniquely to the variables. However, Storkel (2004b) argued that with the strong correlation between ND and WL, such a regression would yield inflated standard errors for the regression coefficients leading to questionable data interpretation. This research puts this hypothesis to the test in a language not previously study with respect to the influence of statistical characteristics on word learning. At present how much variance in vocabulary size is accounted for by ND, WF and WL is not known, and consequently collinearity among the variables has not yet been explored. Finally, before turning to the analysis, the learning heuristics alluded to above are introduced.

Possible learning heuristics

Two speculative hypotheses were generated to explain the results for English and French (Stokes, 2010, Stokes et al., 2011). First, there is evidence that adult speakers implicitly regulate high density words to expand vowel space and increase vowel duration, possibly to reduce listener confusion among high density words (e.g., Munson, 2004; Scarborough, 2004; Wright, 2004). It is possible that these high density words become more salient to children at the early stages of lexical emergence as children take advantage of exaggerated cues in the input to break

into word learning. An extensive body of research on infant use of prosodic cues supports this notion (e.g., Thiessen, Hill & Saffran, 2005).

Second, it is possible that dense words, that by definition share lead (CV#, for example hat, ham, have), rhyme (#VC, for example hat, cat, rat) or consonant segments (CVC, e.g. hat, hot, hit) provide a familiar phoneme stream which becomes readily recognizable, and acts to facilitate new word learning by activating a narrow network of words to which new words can be anchored. The resulting impact of this could be that dense words are less taxing of auditory-verbal short-term memory abilities than words from sparse neighborhoods (Saffran & Graf Estes, 2006; Swingley, 2005), because familiar word shapes would be activated for production. Late talkers may learn denser words because of the reduced short-term memory load. This argumentation is congruent with findings that LTs had poorer verbal short-term memory abilities than their TD peers (e.g., Stokes & Klee, 2009a; 2009b). Children who struggle to learn a lexicon may not only employ statistical learning (e.g. Saffran, 2003), but may remain in a phase of *extended statistical learning,* while children with better verbal short-term memory abilities pass through this phase of statistical learning to begin to learn words from sparse neighborhoods. This suggestion, that failure to abandon a successful strategy in favor of another inhibits further learning, is not entirely new. Aslin and Newport (2008) suggested that maintaining an early effective constrained statistical learning mechanism could 'block' later learning in some children.

The intention in the current research is not to directly explore these hypothesized learning heuristics, but to approach the hypotheses indirectly. This is achieved by exploring the relative value of ND, WF and WL as predictors of vocabulary size in Danish. The research questions were:

1. How much variance in vocabulary size is accounted for by neighborhood density (ND), word frequency (WF) and word length (WL) together and independently in Danish-speaking two-year-old children? The hypothesis is that ND will account for little of the variance in vocabulary scores in Danish and WF and/or WL will be stronger cues, due to the phonological properties of the language.

2. Is there a significant difference in ND, WF and WL among groups of children defined by low, average, and high vocabulary size? An A-prori hypothesis for Danish was not generated.

Method

Participants

The original Danish sample consisted of 922 children (477 girls) aged between 26 and 30 months (M = 27.94, SD = 1.43). The children were sourced from a group of 6,112 children (age range 8 to 36 months) recruited for the norming of the Danish CDI (Bleses, Vach, Slott, Wehberg, Thomsen, Madsen & Basbøll, 2008b). The children were selected randomly from the Danish population register. The inclusion criteria were aged between 26 and 30 months, monolingual Danish children living with both parents and having no reported speech, hearing or other serious (chronic) health problems (note that the occurrence of otitis media followed by tympanostomy tube or premature birth at a gestational age of 32 weeks or later did not lead to exclusion). The Danish project was registered with the Danish Data Protection Agency, as required.

Materials and procedure

The construction of the Danish Communicative Developmental Inventories (DCDI) is reported in detail in Bleses, et al (2008b). The checklist consists of 725 items. The DCDI was mailed to parents and the parents completed the forms and mailed them back directly to the

research group. On the MCDI the parent checks off each word that they know is known by their child. In this study, words checked by the parent as understood and spoken were entered into an SPSS database. Only monosyllabic words were considered, for two reasons. First, this is in line with previous research (e.g., Storkel, 2004a, Zamuner, 2008; Stokes, 2010). Second, while most monosyllabic words have some phonological neighbors, and some words longer than one syllable do too (examples from English are converse, converge and convert), many do not (e.g., popcorn) and including bi- or multi-syllabic words would seriously bias results. In line with prior research (Stokes, 2010), words from the following DCDI categories were excluded: sound effects and animal sounds, people, games and routines, words about time, pronouns, questions words, prepositions and locations, quantifiers and articles, helping verbs and connecting words. This was to centre the analysis on core vocabulary rather than words likely to be context based ('people') or function words. There were 146 Danish words, 104 nouns, (71%), 7 verbs (4%) and 36 adjectives (24%).

Neighborhood density (ND), word frequency (WF) and word length (WL).

A Danish corpus of child directed speech was used to derive ND and WF values for this study. The Danish corpus consists of 51,620 utterances containing a total of 228,661 (coded) words. The words were derived from two corpora: the Odense Twin Corpus (Basbøll, Bleses, Cadierno, Jensen, Ladegaard, Madsen & Thomsen, 2002) and the Plunkett corpus (Plunkett, 1985, 1986). Words were transcribed phonologically in the OLAM database (Madsen, Basbøll, & Lambertsen, 2002). ND values were derived using the Luce and Pisoni (1998) metric (+/- one phoneme substitution, addition or deletion). WF was defined as the number of times that a given word occurs in the corpus. WL was the number of phonemes in the word. Each word checked off

by a parent as used (spoken) by the child was coded for ND, WF and WL, and mean ND, WF and WL values were generated for each child.

Results

Data distributions and initial data reduction

Children with fewer than 20 core words were deleted from the analysis, as had been done for English, leaving data from 894 Danish children in the analysis. The mean DCDI, ND, WF, WL and age in months are shown in Table 2.

Table 2 about here

ND, WF and WL as predictors of vocabulary size (DCDI)

The first research question was 'How much variance in vocabulary size is accounted for by neighborhood density (ND), word frequency (WF) and word length (WL) together and independently in Danish-speaking two-year-old children?' The predictor variables (ND, WF and WF) had small but significant correlations with age (Table 3). More striking were the moderate and high significant correlations (some negative, some positive) among the predictor variables. These results suggested that multicollinearity would be problematic in the planned regressions. Separate third-order partial correlations revealed significant relationships between each predictor variable and the outcome variable (vocabulary size; Table 4). The amount of variance in vocabulary size uniquely attributable to ND, WF and WL, as calculated by simple third-order partial correlations, was 25%, 12% and 7% respectively. Therefore all variables were retained in a regression analysis.

Tables 3 and 4 about here

ND, WF, Age and WL were entered as predictors of vocabulary size in a multiple regression using forced entry. Together the variables accounted for 47% of the variance in

vocabulary size. The *t* values suggested that ND accounted for the most variance in vocabulary scores, followed by WF, Age and WL (*t* = -14.83, -8.23, 6.52 and -6.12 respectively, all $p <$ 0.001). Indicators of collinearity diagnostics suggested that the relationships among the predictors were not seriously problematic. The variance inflation factor ratings (VIF) ranged from 1.02 to 2.69 and the tolerance statistic ranged from .37 to .98. The standard errors of the coefficients were small relative to the size of the coefficients (Table 5).

In order to assess combined and unique variance in vocabulary size accounted for, variables were entered into a regression using the forward method. The total model accounted for 47.2% of the variance in vocabulary size (adjusted $R^2 = 0.47$). ND accounted for 39% of unique variance (*F* change (1, 892) = 570.18), WF added an additional 3.2% of unique variance (*F* change (1,891) = 49.93, Age added 2.8% of unique variance accounted for (*F* change (1, 890) = 44.72), and WL added 2.2% of unique variance accounted for (*F* change (1,889) = 37.44), all *p*s < 0.001, Table 5. ND and WF decreased and WL increased as vocabulary size increased. Data were transformed into Z scores for plotting the relationships (Figure 1).

Table 5 and Figure 1 about here.

Comparison of low, average, and high vocabulary groups

Group differences in ND, WF and WL. There second research question was 'Is there a significant difference in ND, WF and WL among groups of children defined by low, average, and high vocabulary size?' Three cut points were used to define low, average, and high vocabulary groups. The first cut point was at or below the 10th percentile for age in months (< -1.26 below the mean Z score), the second was within 1 *SD* of the mean for age in months, and the third was at or above the 90th percentile for age in months (> 1.26 above the mean Z score).

Using these cut points, 110 children were categories as low vocabulary (LV), 602 were categorized as average (AV), and 83 were categorized as high (HV).

The three groups were significantly different on ND (*Chi-Square*; $\chi^2(2) = 141.83$, $p < 0.001$). For all subsequent between-group comparisons, equal variances were not assumed (t-test). The LV group had significantly higher ND values than the AV group ($t(114.56) = 11.44$, $p < 0.001$, mean difference = .84, *CI* = .75 – .92) and the HV group ($t(114.35) = 13.73$, $p < 0.001$, mean difference = 1.00, *CI* = .85 – 1.15). The AV group had significantly higher ND values than the HV group ($t(294.09) = 10.33$, $p < 0.001$, equal variances not assumed, mean difference = 0.16, *CI* = .13 - .19).

The three groups were also significantly different on WF ($\chi^2(2) = 130.29$, $p < 0.001$). The LV group had significantly higher WF values than the AV group ($t(112.10) = 8.42$, $p < 0.001$, mean difference = 38.40, *CI* = 29.36 – 47.43) and the HV group ($t(111.00) = 10.25$, $p < 0.001$, mean difference = 46.60, *CI* = 37.58 – 55.61). The AV also had significantly higher WF scores than the HV group ($t(398.37) = 11.84$, $p < 0.001$, mean difference = 8.19, *CI* = 6.83 – 9.55).

There was also a significant difference between the three groups on WL ($\chi^2(2) = 37.33$, $p < 0.001$). The LV group had significantly lower WL values than the AV group ($t(113.34) = 7.32$, $p < 0.001$, mean difference = .08, *CI* = .05 - .09) and the HV group ($t(112.88) = 7.34$, $p < 0.001$, mean difference = .08, *CI* = .05 - .09). There was no significant difference between the AV and HV group for WL.

Results summary

For the entire sample of 894 children, in a multiple regression ND accounted for 39% of the variance in vocabulary size and WF, Age and WL contributed an additional 3.2%, 2.8% and 2.2% of unique variance accounted for respectively. Children with small lexicons had

significantly higher ND and WF and significantly lower WL than children with average and high vocabulary sizes.

Discussion

The purpose of this research was to determine the cross-linguistic validity of the claim that the lexicons of two-year-old children reflect children's preference for learning words that are comprised of words from dense phonological neighborhoods in the ambient input (e.g., Stokes, 2010; Storkel, 2004a) and that children who are struggling to learn their ambient language are learning words that come from denser neighborhoods than the words of their same-age peers (e.g., Stokes, 2010; Stokes, et al, 2011). This research confirms these findings for a third language, but one important difference emerged. The hypotheses are discussed in turn.

The first hypothesis was that due to the high number of vowels relative to consonants, and the phonotactic qualities of Danish which rendered syllable boundaries indistinct, Danish-speaking two-year-old children may not show evidence of having learned words that came from phonologically dense neighborhoods in the ambient language stream because the same learning opportunity would not be found in Danish as would be found in languages like English and French. This hypothesis was disproved. While the amount of variance in lexicon size accounted for by ND was less than that for English (47%) and French (53%), it was still remarkably high. This finding was unexpected as it was anticipated that fewer ND cues would be available to Danish-speaking children than languages with larger consonant inventories and clearer syllable boundaries. While the hypothesis as stated was disproved, different hypotheses emerge.

As Bleses et al (2010) state, where the vowel:consonant ratio is high, word segmentation is compromised. This may mean that the Danish children are responding to consonant neighbors (words that differ by vowels, for example hot, hat, hut) rather than lead or rhyme neighbors. This

premise is worthy of investigation as it is possible that as the vowel:consonant ratio makes Danish children sensitive to vowel distinctions earlier than other languages and Danish-speaking late talkers may rely more on consonant neighbors than lead or rhyme neighbors. The difference is theoretically important because it could provide clues to how learners make use of variants of statistical cues, like neighborhood density, dependent on the specific structure of the input.

There were no specific hypotheses for word frequency and word length, given that Danish has not previously been examined for these characteristics. There were significant correlations among the predictors but the results of a forced entry regression (that is, with no A-priori assumption of the value of one predictor over another) implied that ND, WF and WL all accounted for a proportion of the variance in vocabulary size and that all variables have an impact on word learning.

Our second question asked if children with small vocabularies (at or below the 10th percentile for age) would show a stronger preference for word from dense neighborhoods than their age-matched peers with average and large vocabularies. Children with small vocabularies had learned, on average, words that came from denser neighborhoods than their peers with larger lexicons. This was confirmed. Although no direct assessment was made of factors linked to this observation, we had hypothesized that this preference for high density words may reflect psycholinguistic processing constraints specific to verbal short-term memory. Other research has shown that children with small lexicons scored significantly lower than children with average to large lexicons on a test of nonword repetition (Stokes & Klee, 2009a; 2009b). Thus poorer verbal short-term memory skills could limit a child's ability to process novel words strings (words from sparse phonological neighborhoods), limiting the likelihood that a child will progress through the early stage of high density as a learning cue. Swingley (2005) convincingly described the

likelihood that emerging lexicons make use of repeated familiar phonotactic strings, as would be experienced in word from dense input neighborhoods. LTs or children with very small lexicons for their age may remain in a stage of *Extended Statistical Learning* longer than their more able peers who progress through this stage to begin to learn words of lower ND in the input. The clear difference between ND scores in the current study for late talkers, typical developers and precocious talkers would seem to support this view, but only a longitudinal study could answer the question definitively.

The data for English, French, and Danish all reflected the same relationship between ND and vocabulary size. An interesting difference emerged for Danish in the direction of the relationship between WF and MCDI scores. WF was positively (directly) related to CDI scores in English and French, and negatively (indirectly) related to CDI scores in Danish, meaning that for English and French small lexicons were comprised of low frequency words, and for Danish, they were comprised of high frequency words. It is possible that this difference can be attributed to the lack of verbs in the Danish dataset. Goodman et al (2008) reported that when a range of word classes is represented in the data under investigation, high frequency words are learned later than low frequency words, as we found for English and French. However, when nouns alone are studied, Goodman et al (2008) found a strong correlation between age of acquisition and word frequency (r = .55), meaning that high frequency words were learned before low frequency words. This same finding occurred for the current Danish data explored here, where there were many more nouns than verbs or adjectives (71%, 4% and 24% respectively) and the partial correlation between WF and vocabulary size, controlling for age, was r = -.54. Very small lexicons were comprised of high frequency words, unlike the English and French studies, and as the lexicon expanded, word frequency decreased. The English and French datasets had a

greater percentage of verbs, 31.5% and 25% respectively. Since the basic form of Danish verbs is the infinitive, which is formed by a syllabic suffix, Danish verbs nearly all are polysyllabic, reducing heavily the number of monosyllabic verbs available for analysis. The result for the Danish word set is congruent with prior findings of a strong positive (direct) relationship between vocabulary size and word frequency for English (e.g., Storkel, 2004a).

Study limitations

While much work has shown that lexical and sublexical distributional properties do influence word learning in English, little work has yet emerged on other languages and there are few, if any, cross-linguistic comparisons. As with all new inroads in science, it is probable that the methods employed here could be improved upon and further cross-linguistic reports will appear. As with other studies of this type (Stokes, 2010, Stokes et al, 2011), this work is limited by a) the lack of consideration of other cognitive, social, linguistic and perceptual factors that also impact on word learning, b) the inclusion of only monosyllabic words, and c) consideration of only expressive, and not receptive, lexicons.

Implications

This line of research has raised an interesting implication for the study of emerging lexicons. Languages of differing phonological properties may appear to present different challenges for young children attempting to learn a lexicon, however, for English, French and Danish children, small lexicons were comprised of words from dense neighborhoods in the ambient input. Whether or not children who continue to struggle to expand their lexicons into the third year of life continue to use this ND strategy remains to be seen. At present, the extended statistical learning hypothesis is just that, a hypothesis. Also, the hypothesis that there is a strong

link between verbal short-term memory skills and use of density as a learning mechanism remains a hypothesis, worthy of exploration because of the implied links outlined above.

Further research is also required to explore cross-linguistic differences in the rate of learning of nouns and verbs because an intriguing question arises from considering the outcomes of the current study and that of Li and Zhao (2007). If Chinese and English verbs and nouns differ in their relative word lengths, do they also differ in ND and/or WF and how might that contribute to the cross-language differences already identified in the proportion of nouns and verbs in emerging lexicons?

References

Aslin, R. N., & Newport, E. L. (2008). What statistical learning can and can't tell us about language acquisition. In J. Colombo, P. McCardle & L. Freund (eds.), *Infant pathways to language: Methods, models and research directions* (pp. 15-29). Hove: Psychology Press.

Basbøll, H. (2005). *The phonology of Danish*. Oxford: Oxford University Press.

Basbøll, H. (2006). Syllabic and morphological structure: what can be learnt from their interaction in Danish? *Working Papers in Language Acquisition, 3*.

Basbøll, H., Bleses, D., Cadierno, T., Jensen, A., Ladegaard, H. J., Madsen, T. O, & Thomsen, P. (2002). The Odense language acquisition project. *Child Language Bulletin, 22*, 11-12.

Bishop, D. V. M., Price, T. S., Dale, P. S., & Plomin, R. (2003). Outcomes of early language delay: II. Etiology of transient and persistent language difficulties. *Journal of Speech, Language, and Hearing Research, 46*, 561-575.

Bleses, D. Basbøll, H., Lum, J., & Vach (2010). Phonology and lexicon in a cross-linguistic. perspective: the importance of phonetics – a commentary on Stoel-Gammon's 'Relationships between lexical and phonological development in young children. *Journal of Child Language*. 38, 1, s. 61-68.

Bleses, D., Basbøll, H., & Vach, W. (on-line 2011): Is Danish difficult to acquire? Evidence from Nordic past tense studies, *Language and Cognitive Processes*.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008a). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language, 35*, 619-650.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008b). The Danish Communicative Developmental Inventories: validity and main

developmental trends. *Journal of Child Language, 35*, 651-669.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2004). Linguistic constraints on statistical computations: the role of consonants and vowels in continuous speech processing. *Psychological Science, 16*, 451-459.

Carreiras, M., & Price, C. J. (2008). Brain activation for consonants and vowels. *Cerebral Cortex, 18*, 1727-1735.

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *International Journal of Language and Communication Disorders, 43*, 361-389.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual.* Baltimore: Brookes.

Fenson, L., Marchman, V. A., Thal, D., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur- Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Baltimore: Brookes.

Fourgeron, C. & Smith, C. L. (1999). French. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press. pp. 78-81.

Goodman, J. C., Dale, P. S., & Ping, L. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35,* 515-531.

Grønnum, N. (2003). Why are the Danes so hard to understand? In H. Galberg Jacobsen, D. Bleses, T. O. Madsen & P. Thomsen (Eds), *Take Danish – for instance: Linguistic studies in honour of Hans Basbøll presented on the occasion of his 60th birthday 12 July 2003,* 119-30. Odense: University Press of Southern Denmark.

Haspelmath, M., Dryer, M. S., Gil, D., B. & Comrie, B. (eds.), (2008). *The world atlas of language structures online*. Munich: Max Planck Digital Library. Available online at http://wals.info/feature/1. Accessed on 21st January, 2010.

Hieronymus, J. L. (1993). *ASCII phonetic symbols for the world's languages: Worldbet*. ASCII Phonetic Symbols for the Worlds Languages: Worldbet Technical report, AT&T Bell Laboratories (1993).

Ladefoged, P. (1999). American English. Handbook of the International Phonetic Association (pp. 41-44). Cambridge: Cambridge University Press.

Li, P., & Zhao, X. (2007). Vocabulary development in English and Chinese: A comparative study with self-organizing neural networks.

Luce, P.A., & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1-36.

Maddieson, I. (2008). Consonant-vowel ratio. In M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie (eds.), Chapter 3. Available online at http://wals.info/feature/1. Accessed on 21st January, 2010.

Madsen, T. O., Basbøll, H., & Lambertsen, C. (2002). OLAM-et semiautomatisk morfologisk og lydstrukturelt kodningssystem for dansk. *Odense Working papers in Language and Communication, 24*, 43-56.

Maekawa, J., & Storkel, H. L. (2006). Individual differences in the influence of phonological characteristics on expressive vocabulary development by young children. *Journal of Child Language, 33*, 439–459.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighbood density on vowel articulation. *Journal of Speech, Language, and Hearing Research, 47,* 1048-1058.

New, B., Brysbaert, M., Veronis, J. & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. Applied Psycholinguistics, 28, 661-677.

Paul, R. (1996). Clinical implications of the natural history of slow expressive language development. *American Journal of Speech-Language Pathology, 5*, 5-21.

Plunkett, K. (1985). *Preliminary approaches to language development*. Århus: Århus University Press.

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

Rischel, J. (1970). Consonant gradation: A problem in Danish phonology and morphology. In H. Benediktsson (Ed.), *Proceedings of the international conference on Nordic and General Linguistics*, University of Iceland, Reykjavík (pp. 460_480). Reykjavík: Vísindafélag Íslendinga. (Reprinted in *Sound structure in language,* pp. 26-43, by N. Grønnum, F. Gregersen, & H. Basbøll, Eds., 2009, Oxford: Oxford University Press).

Rischel, J. (2003). The Danish syllable as a national heritage. In H. Galberg Jacobsen, D. Bleses, T. O. Madsen & P. Thomsen (eds), *Take Danish – for instance : linguistic studies in honour of Hans Basbøll presented on the occasion of his 60th birthday* 12 July 2003, 273–282. Odense: University Press of Southern Denmark.

Saffran, J. R. (2003).Statistical language learning: Mechanisms and Constraints. *Current Directions in Psychological Science, 12*, 110-114.

Saffran, J. R., & Graf Estes, K. (2006). Mapping sound to meaning: Connections between learning about sounds and learning about words. *Advances in Child Development and Behavior, 34*, 1-38.

Scarborough, R. A. (2004). *Coarticulation and the structure of the lexicon*. Unpublished

doctoral dissertation, University of California, Los Angeles. Accessed on 7th May, 2009, from http://www.linguistics.ucla.edu/faciliti/research/scarb_diss.pdf

Stokes, S. F. (2010). Neighborhood density and word frequency in toddlers. *Journal of Speech, Language, and Hearing Research, 53*, 670-683.

Stokes, S. F., Kern, S., & dos Santos, C. (2011). Extended statistical learning as an account for slow vocabulary growth. *Journal of Child Language.*

Stokes, S. F., & Klee, T. (2009a). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry, 50*, 498-505.

Stokes, S. F., & Klee, T. (2009b). The diagnostic accuracy of a new test of early nonword repetition for differentiating late talking and typically developing children. *Journal of Speech, Language, & Hearing Research, 52*, 872-882.

Storkel, H. L. (2004a). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25*, 201-221.

Storkel, H. L. (2004b). The emerging lexicon of children with phonological delays. *Journal of Speech, Language, and Hearing Research, 47*, 1194-1212.

Storkel, H. L. (2008a). First utterances. In G. Rickheit and H. Strohner, (eds.), *Handbook of communication competence* (pp. 125-147). Berlin: Mouten de Gruyter.

Storkel, H. L. (2008b). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language, 36,* 291-321.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86-132.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*, 53-71.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in Laboratory Phonology VI* (pp. 75-87). Cambridge: Cambridge University Press.

Zamuner, T. S. (2008). The structure and nature of phonological neighborhoods in children's early lexicons. *Journal of Child Language, 36*, 3-21.

Table 1

Selected Phonological Features of English, French and Danish.

Feature	English[a,b]	French[c]	Danish[g]
Stress	Stress-timed	Syllable-timed	Stress-timed
Rhythm	Trochaic	Undetermined	Trochaic
Consonants[d]	24 (Average)	21 (Average)	18 (Average)
Vowels[d]	15 (Large)	14 (Large)	37 (Very Large)
C/V Ratio[e]	1.6 (Low)	1.5 (Low)	.49 (Very Low)
Phonotactics	C(0-3)VC(0-4) [f]	C(0-3)VC(0-3)	C(0-3)VC(0-3)
Monosyllabic (%)	75.69	72.62	55.12
Bisyllabic (%)	16.63	20.67	38.97
Open syllables (%)	27.48	55.71	32.64
Closed syllables (%)	55.71	26.09	53.90
V only syllables (%)	16.81	17.22	13.45

Note. [a]Ladefoged (1999). [b]Haspelmath, Dryer, Gil & Comrie (2008). [c]Fourgeron & Smith (1999).[d]Inventory size and categorization (small, large etc.) for consonants and vowels, where vowels includes diphthongs, size notation is from Maddieson (2008). [e]C/V = Consonant/Vowel Ratio, a representation of the phonological complexity of languages (Maddieson, 2008). [f]The notation (0-4) indicates that the number of final consonants could be 0, 1, 2, 3, or 4. Derived from Lexique3 (New et al., 2007). [g]Hieronymus (1993).

Table 2

Means (Standard Deviations) and Range Scores for Key Variables for Danish-speaking Children (N = 894).

	Mean (SD)	Range
Vocabulary size	424.68 (144.61)	57 - 725
Age (months)	27.94 (1.43)	26 - 30
Neighborhood density (ND)	10.96 (.46)	10.11 – 13.89
Word frequency (WF)	99.89 (24.29)	44 – 288
Word length (WL)	3.30 (0.6)	2.96 - 3.46

Table 3

Correlations Between the Predictor Variables and the Outcome Variable.

Correlations

		ND	WF	WL	month	totalDCDI
ND	Pearson Correlation	1	.655**	-.737**	-.130**	-.624**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	894	894	894	894	894
WF	Pearson Correlation	.655**	1	-.572**	-.098**	-.545**
	Sig. (2-tailed)	.000		.000	.003	.000
	N	894	894	894	894	894
WL	Pearson Correlation	-.737**	-.572**	1	.066*	.377**
	Sig. (2-tailed)	.000	.000		.050	.000
	N	894	894	894	894	894
month	Pearson Correlation	-.130**	-.098**	.066*	1	.249**
	Sig. (2-tailed)	.000	.003	.050		.000
	N	894	894	894	894	894
totalDCDI	Pearson Correlation	-.624**	-.545**	.377**	.249**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	894	894	894	894	894

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 4

Third Order Partial Correlations Between Predictors and Outcome Variables (Danish).

Predictor variable	Correlation with DCDI (df)	Controlled variables
ND	-.50, $p < 0.001$ (889)	Age, WF, WL
WF	-.35, $p < 0.001$ (889)	Age, WL, ND
WL	-.26, $p < 0.001$ (889)	Age, WF, ND

Table 5

Table of Coefficients for Danish (Predicting Vocabulary Size).

Model	*B*	*SE B*	β	*t*	*CI* lower bound	*CI* upper bound
1						
(Constant)	2555.96	89.34	-.62	28.61	2380.63	2731.29
ND	-194.45	8.14		-23.88	-210.43	-178.47
2						
(Constant)	2165.41	103.05		21.01	1963.15	2367.67
ND	-145.90	10.49	-0.46	-13.90	-166.49	-125.31
WF	-1.41	0.20	-0.23	-7.06	-1.81	-1.02
3						
(Constant)	1621.99	129.34		12.54	1368.15	1875.83
ND	-139.91	10.28	-.50	-13.60	-160.09	-119.73
WF	-1.39	.20	-0.23	-7.12	-1.77	-1.01
Age	17.01	2.55	.17	6.68	12.02	22.01

Table continues

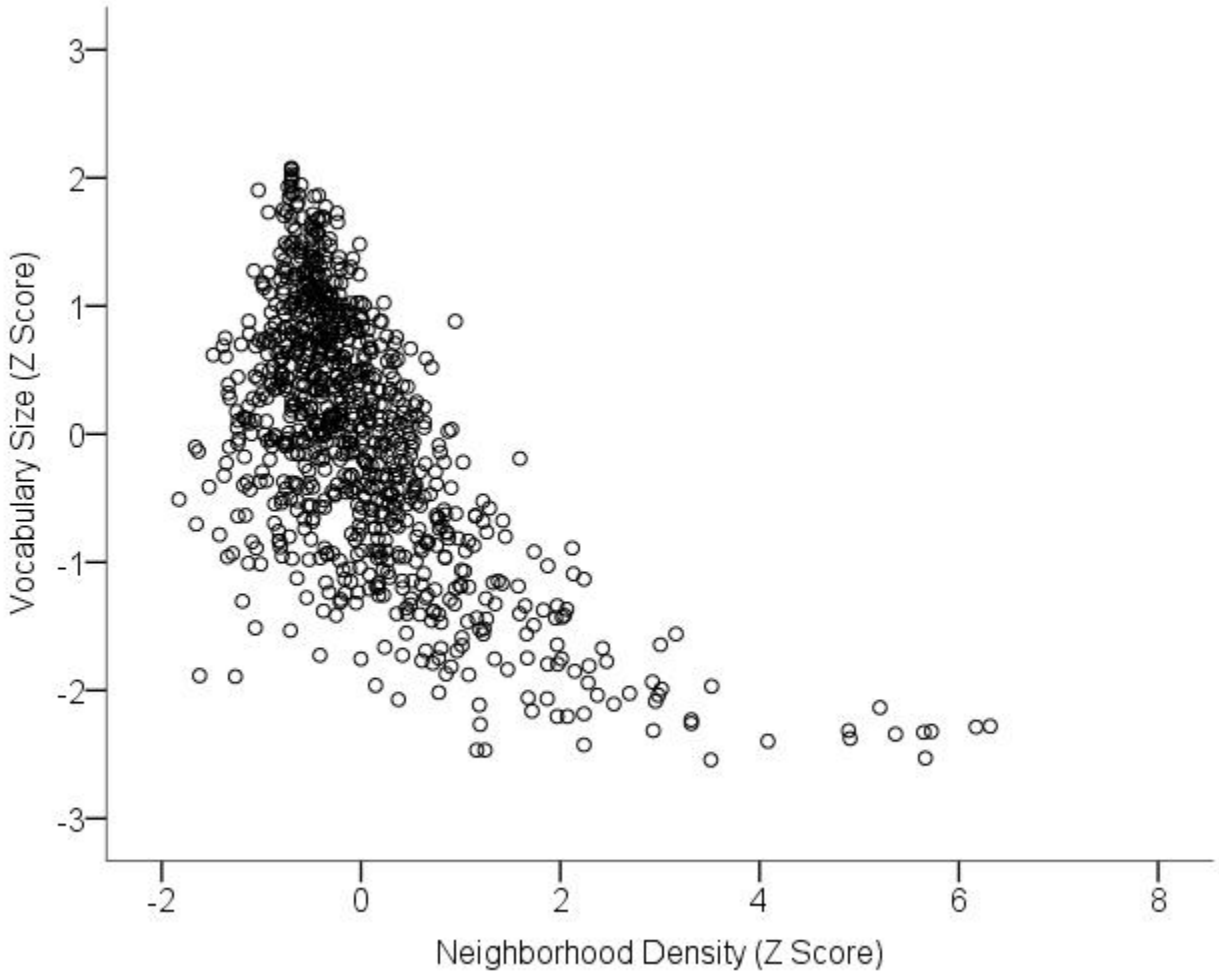
Model	*B*	*SE B*	β	*t*	*CI* lower bound	*CI* upper bound
4						
(Constant)	4096.993	423.90		9.66	3265.04	4928.94
ND	-184.559	12.44	-.59	-14.83	-208.97	-160.14
WF	-1.61	.20	-.27	-8.22	-1.98	-1.22
Age	16.277	2.50	.16	6.51	11.37	21.17
WL	-588.806	96.23	-.22	-6.11	-777.67	-399.94

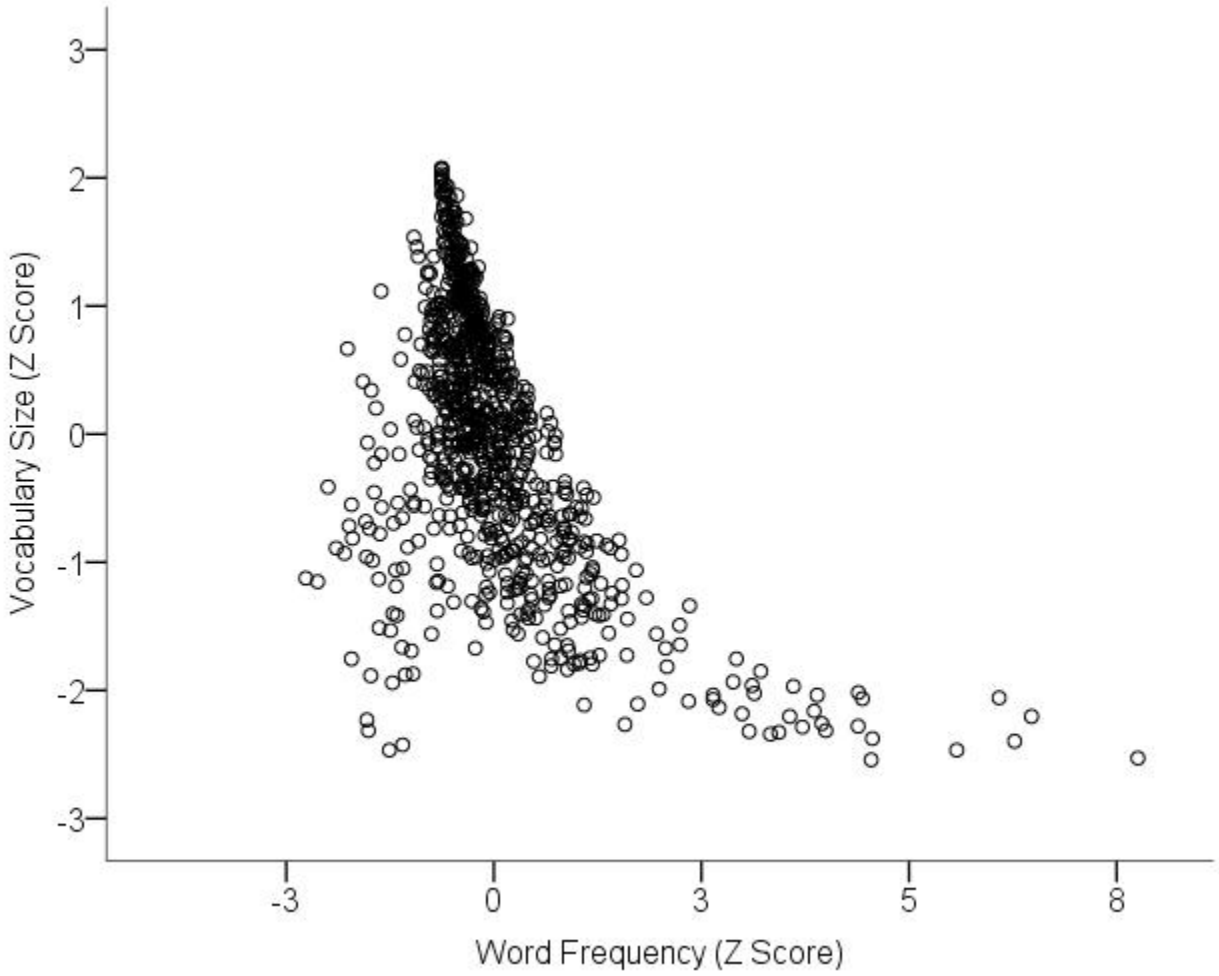
Note: All *p*s < 0.001, ND = neighborhood density, WF = word frequency, Age = age in months, WL = word length.

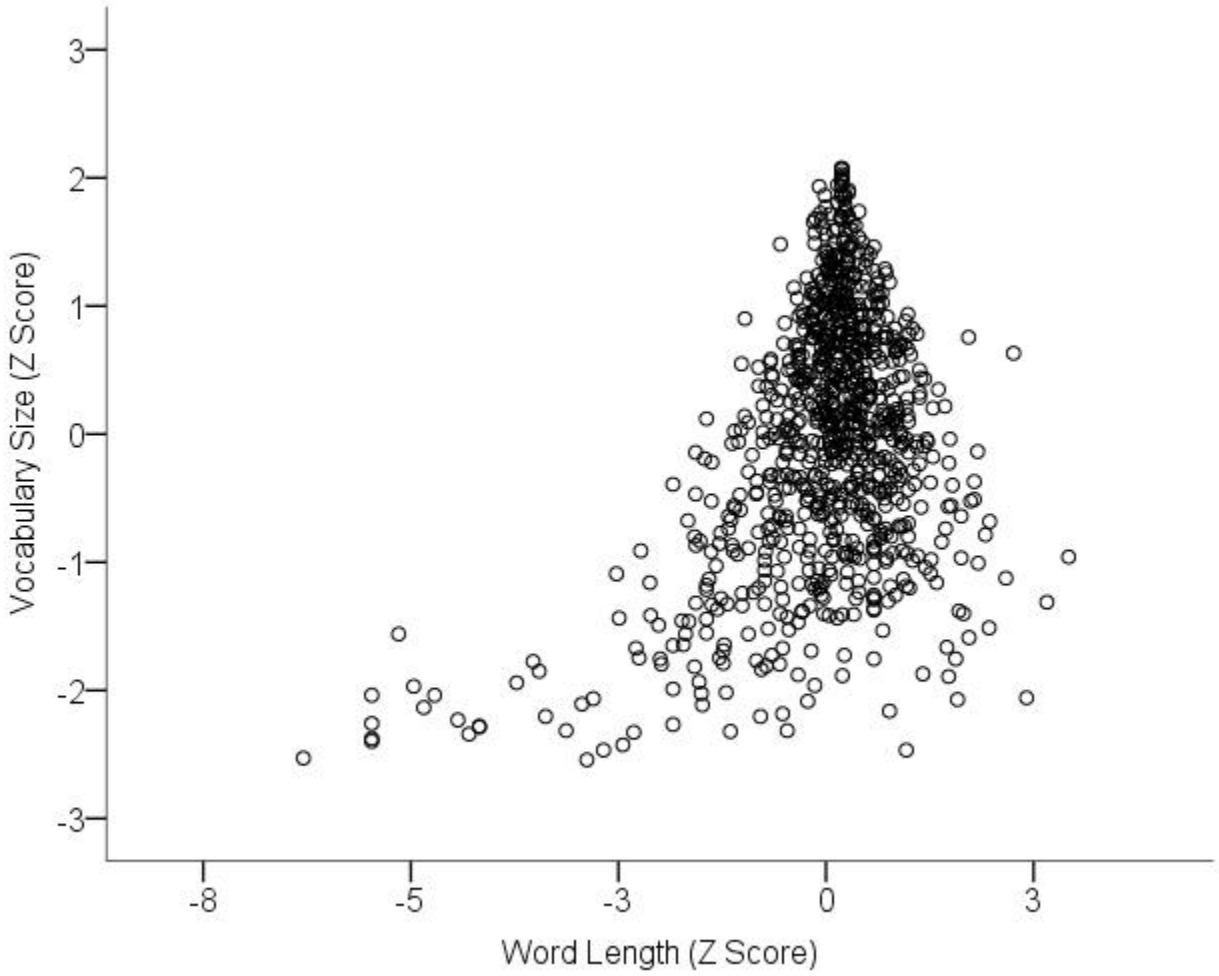
Figure 1. Scatterplot of the relationship between vocabulary size and neighborhood density for Danish.

Figure 2. Scatterplot of the relationship between vocabulary size and word frequency for Danish.

Figure 3. Scatterplot of the relationship between vocabulary size and word length for Danish.


Vocabulary Size (Z Score)
3
2
1
0
-1
-2
-3
-2
0
2
4
6
8
Neighborhood Density (Z Score)


Vocabulary Size (Z Score)
3
2
1
0
-1
-2
-3
-3
0
3
5
8
Word Frequency (Z Score)


Vocabulary Size (Z Score)
3
2
1
0
-1
-2
-3
-8
-5
-3
0
3
Word Length (Z Score)