J. Child Lang. **36** (2009), 629–656. © 2008 Cambridge University Press
doi:10.1017/S0305000908009150X Printed in the United Kingdom

Prosodic patterns in Hebrew child-directed speech*

OSNAT SEGAL, BRACHA NIR-SAGIV,
LIAT KISHON-RABIN AND DORIT RAVID
Tel-Aviv University

(*Received* 25 *December* 2006. *Revised* 9 *September* 2007.
First published online 13 *November* 2008)

ABSTRACT

The study examines prosodic characteristics of Hebrew speech directed to children between 0;9–3;0 years, based on longitudinal samples of 228,946 tokens (8,075 types). The distribution of prosodic patterns – the number of syllables and stress patterns – is analyzed across three lexical categories, distinguishing not only between open- and closed-class items, but also between these two categories and a third, innovative, class, referred to as between-class items. Results indicate that Hebrew CDS consists mainly of mono- and bisyllabic words, with differences between lexical categories; and that the most common stress pattern is word-final, with parallel distributions found for all categories. Additional analyses showed that verbs take word-final stress, but nouns are both trochaic and iambic. Finally, a developmental analysis indicates a significant increase in the number of iambic words in CDS. These findings have clear implications regarding the use of prosody for word segmentation and assignment of lexical class in infancy.

One critical task an infant faces is to segment the oncoming stream of speech in order to determine word boundaries and detect the sequence of sounds that represent words in his ambient language. An apparent difference between adult language users and infants is that the latter are unlikely to reliably use lexical knowledge in order to determine word boundaries. Instead, recognizing sublexical regularities based on their relative weighting in the language – and particularly regularities of prosodic cues – is considered a powerful procedure for segmentation and language

[*] This work was performed in partial fulfillment of the requirements for a PhD of Osnat Segal, Sackler Faculty of Medicine, Tel Aviv University, Israel. We are indebted to three anonymous reviewers and to the associate editor of the journal for their detailed and helpful comments. The authors are alone responsible for any inadequacies that remain. Address for correspondence: Osnat Segal, 7b Zelig Bas, Petach Tiqwa, Israel. e-mail: segalll@netvision.net.il

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

acquisition in infancy (Jusczyk, Houston & Newsome, 1999; Mattys & Jusczyk, 2001). It has well been established that word stress is an important cue for segmentation in adults and infants (e.g. Cutler & Norris, 1988; Jusczyk, Houston & Newsome, 1999). However, stress-based segmentation has usually been documented in trochaic languages, in which stress is assigned to the first syllable. Thus, the generalizations related to the results of such studies could be questioned: it is not clear whether the use of prosodic patterns for segmentation is influenced by the typological characteristics of the languages in question (i.e. trochaic) or possibly by a universal, auditory perceptual bias for stress-initial segmentation (Echols & Newport, 1992; Echols, Crowhurst & Childers, 1997). A major question is thus whether language-specific regularities of stress patterns affect infants from early on, or whether they are predisposed to universal prosodic cues such as initial stress. One way to resolve this issue is by studying the development of segmentation procedures in iambic languages as a counterpoint to the extensive study of, for example, English as a representative of trochaic languages. The study of iambic languages allows for a distinction between preferences that are based on the most frequent prosodic patterns and the effects of a possible auditory bias such as stress-initial segmentation. This line of research requires first a detailed description of the characteristics of the prosodic input.

To date, the few studies on segmentation in iambic languages focus mainly on data from French. However, French is typically considered a syllable-timed rather than a stress-timed language with possible final lengthening (Nazzi, Iakimova, Bertoncini, Frédonie & Alcantara, 2006). The analysis of Hebrew as a clearly iambic, although not a syllable-timed language, may thus prove more comparable to the results documented for English.

The current paper focuses on mapping the common prosodic patterns in Hebrew, based on a database of child-directed speech. Our goal is to provide a detailed analysis of the prosodic patterns in Hebrew, as well as their regularities in spoken language, in order to offer a basis for interpreting future results of studies on segmentation and speech perception in infants. We start by describing the problem of segmentation in early language development, while considering the role of prosodic cues such as stress and word length. We than proceed to explain the importance of studying prosodic patterns and their regularities in iambic languages, and especially in Hebrew.

The problem of segmentation

The role of prosody in speech processing and language development has been investigated extensively during the last decades. The prosodic patterns

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

of a word – for example, the number of syllables and primary stress – were found to provide important cues in word recognition tasks performed by adults (Arciuli & Cupples, 2004) and in word segmentation tasks performed by both adults (Cutler & Norris, 1988) and infants (Echols *et al.*, 1997; Jusczyk, Houston & Newsome, 1999; Nazzi, Dilley, Jusczyk, Shattuck-Hufnagel & Jusczyk, 2005). The development of segmentation procedures is critical for young language learners in building their vocabulary from fluent, unfamiliar speech because there are no reliable pauses between words and there is no one systematic acoustic marking of word boundaries (Cole & Jakimik, 1980). Adults may rely on top-down procedures when segmenting speech, using stored knowledge of morpho-phonological structures and matching them with sequences from the speech stream for the extraction of word boundaries. Infants, however, first need to be able to segment the speech stream into words, and so should construct this knowledge by relying on bottom-up strategies that involve acoustic–prosodic cues (e.g. Jusczyk, Houston & Newsome, 1999).

To make things even more complicated, words in isolation are rarely represented in speech directed to children. The bulk of what infants hear consists of words embedded in utterances, even in situations where caregivers are encouraged to teach their infants new words. These words are presented in the context of utterances most of the time (Woodward & Aslin, 1990), and approximately 90% of all utterances are in the form of continuous speech (Brent & Siskind, 2001). As a result, several researchers have proposed a variety of procedures and cues that may allow infants to identify word boundaries without necessarily 'knowing' the actual words that these boundaries delimit. For example, some of the cues that have been found to be exploited by infants for word segmentation are distributional cues (the odds that one syllable would follow another; e.g. Saffran, Aslin & Newport, 1996), allophonic cues (the realization of some phoneme according to their position in the word; e.g. Jusczyk, Hohne & Bauman, 1999), phonotactic cues (limitations on phoneme order within words; e.g. Mattys & Jusczyk, 2001), coarticulation cues (acoustic influences of adjacent segments; e.g. Johnson & Jusczyk, 2001) and prosodic cues (the detection of stressed syllables and phrasal boundaries; e.g. Jusczyk, Houston & Newsome, 1999; Gout, Christophe & Morgan, 2004).

Prosodic bootstrapping and syllabic stress

The use of prosodic cues to initiate the acquisition of lexicon and syntax has been termed PROSODIC BOOTSTRAPPING (e.g. Gleitman & Wanner, 1982). Prosodic bootstrapping procedures involve tuning into a combination of acoustic and prosodic cues, such as vowel duration, amplitude, pitch, number of syllables, and phonological syllable structure designated for

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

word segmentation and the analysis of grammar throughout language development (e.g. Gout *et al.*, 2004; Shi, Morgan & Allopenna, 1998; Shi, Werker & Morgan, 1999).

One of the main procedures for word segmentation in English and related trochaic languages is relying on syllabic stress (Houston, Jusczyk, Kuijpers, Coolen & Cutler, 2000; Jusczyk, Houston & Newsome, 1999). Syllabic stress refers to the perceptual impression of the listener that one syllable is stronger or more prominent than the other. This impression, however, is typically guided by acoustic cues such as amplitude, fundamental frequency or syllable duration, or by a combination of the three (Lehiste, 1970). The alternation of stressed and unstressed syllables within a word creates a gross amplitude pattern and a prosodic structure that is termed 'word stress pattern' (Jusczyk, Cutler & Redanz, 1993). For example, in bisyllabic words, either the first syllable carries stress, resulting in a strong–weak stress pattern (trochee), or alternatively, the second syllable may carry the stress, yielding a weak–strong stress pattern (iamb).

Polysyllabic words distinguish between primary and secondary stress, based on differences in acoustic cues. Specifically, longer duration and higher amplitude and pitch peak are expected for syllables carrying primary stress, thus creating a rhythm which makes them more prominent for the listener than syllables with secondary (or non-primary) stress (Mattys, 2000).

These features provide young infants with a powerful tool for detecting the rhythmic structure of fluent speech. Specifically, infants are attuned to the rhythmic characteristics of utterances in their language from a very early age. Even neonates display sensitivity both to the prosodic distinctiveness of their native language and to differences between utterances of different languages (e.g. Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988). Furthermore, during the second half of their first year of life (but not before), infants seem to develop a preference for the PREDOMINANT stress pattern of words in their language. Work on English indicates a preference for strong–weak (trochaic) pattern, demonstrated by longer listening time to lists of strong–weak words as compared with lists of weak–strong (iambic) words (Jusczyk *et al.*, 1993). This sensitivity to the typical prosodic shape of English words appears to influence infants' segmentation procedure for bisyllabic words.

In a landmark study, Jusczyk, Houston & Newsome (1999) found that when English-learning infants aged 0;7·15 were presented first with isolated strong–weak (trochaic) words (e.g. *kíngdom, hámlet, dóctor, cándle*), or with weak–strong (iambic) words (e.g. *guitár, devíce, berét, surpríse*) they later showed a preference only to the strong–weak words when presented in passages. Also, when infants were initially presented only with the stressed syllables of iambic words (but not of trochaic words), they later preferred

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

passages that included these stressed segments positioned within iambic words. These results suggest that infants aged 0;7·15 tend to segment speech into units that begin with strong syllables. Furthermore, in the familiarization stage, when presented with passages containing weak–strong words followed consistently by the same function words (e.g. *guitár is*, *surpríse in*, *berét on*, *devíce to*) infants showed a listening preference for the strong–weak pseudo-words (e.g. *táris*, *rísein, réton, víceto*). These results suggest that infants segmented strong–weak units from fluent speech even when these units crossed word boundaries. Thus, when consistent function words followed the bisyllabic words, infants seemed to use the stressed syllable as a cue for locating the beginning of words and the distributional cue of the repeating last weak syllable as a cue for locating their ending, fusing the function word as the second unstressed syllable. The segmentation of weak–strong words (nouns) (e.g. *guitár*) appeared at age 0;10·15. It was concluded that at the first stage of word segmentation, English-learning infants use their sensitivity to the predominant strong–weak stress pattern of English words and treat strong syllables as potential markers of word onsets (see, also, Echols *et al*., 1997).

The findings that infants use the stressed syllable as a cue for word beginning (Jusczyk, Houston & Newsome, 1999) comply with the Metrical Segmentation Strategy (MSS) that was found for adult listeners. The MSS suggests that English-speaking adults posit a word-onset boundary at strong syllables containing full vowels as a first pass strategy in segmenting words from fluent speech (e.g. Cutler & Norris, 1988). However, there are findings indicating that stress-based segmentation is driven by both vowel quality and non-segmental correlates of stress (Spitzer, Liss & Mattys, 2007). The MSS strategy is considered extremely useful for tapping into the lexicon and syntax of a trochaic language such as English, since most content words in English conversational speech (about 90%) begin with a strong syllable (Cutler & Carter, 1987). In fact, English-acquiring infants are so influenced by the frequent strong–weak stress pattern in their language, that at age 1;1·15 they show segmentation advantage to verbs with a strong–weak stress pattern in comparison to verbs with a weak–strong stress-pattern (Nazzi *et al*., 2005) even though most verbs in English (69%) are weak–strong (Kelly & Bock, 1988).

It is important to note that prosodic cues such as word length and stress pattern seem to help young language learners not only to segment words, but also to differentiate between word classes (open-class, or lexical items, vs. closed-class, or functional, grammatical items) as well as differentiating between different subclasses, such as verbs versus nouns (e.g. Kelly & Bock, 1988; Shi *et al*., 1999). For example, grammatical items are shorter in length when counted by number of syllables; they manifest shorter vowel duration, weaker amplitude and simplified syllabic structure (e.g. Shi *et al*.,

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

1998). Christophe, Guasti, Nespor, Dupoux & Ooyen (1997) suggest that function words tend to occur at the borders of phonological phrases – that is, one or two content words together with some function words, possessing one melodic contour – and thus are marked by changes in duration, pitch or loudness of segments in the word. Infants may use this acoustic information to segment and store repeating function items that tend to appear regularly at the end or the beginning of phonological phrases. Recognizing the form of function words can further help infants in stripping these frequent syllables off and detecting the acoustic stress regularity of content words as a tool for segmentation and for constructing a vocabulary.

As for verbs and nouns, Kelly & Bock (1988) reported that in a representative sample of bisyllable English words (3,000 nouns and 1,000 verbs) most strong–weak words (90%) were nouns and most weak–strong words (85%) were verbs. This type of prosodic diversity might influence both the learning of grammatical category assignment (Kelly & Bock, 1988) as well as the development of segmentation strategies during language development (Nazzi *et al.*, 2005). Type versus token frequencies of items belonging to different word classes were also found to provide perceptual cues for the language learner, since grammatical items – a typically limited group in terms of number of types – tend to occur at higher token frequencies compared with lexical items (see also Cutler & Carter, 1987; Shi *et al.*, 1999).

Previous work on iambic languages

The use of stress patterns as a strategy for word segmentation in infants has been primarily investigated in trochaic languages such as English. In iambic languages, limited data exist with the exception of two studies conducted with Canadian-French and European-French-learning infants (Polka & Sundara, 2003; Nazzi *et al.*, 2006). These studies differ in their results although both emphasize the importance of the common rhythmic pattern in the language for initial segmentation procedures. Polka & Sundara (2003) reported that Canadian-French infants aged 0;8 could segment weak–strong words from fluent Canadian-French speech, but did not segment English strong–weak words from English passages. These results supported the hypothesis that segmentation abilities are influenced by the common prosodic structure of the native language. Nazzi *et al.* (2006), however, found that for European-French-learning infants, segmentation of weak–strong words appears between 1;0 and 1;4. At 1;0, infants segmented the iambic words into separate syllables, suggesting that the syllable (whether stressed or not) is the initial unit for segmentation in the iambic European-French. They also note that the difference in findings between these two studies is attributed mainly to the prosodic differences between

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

the two French dialects. Specifically, there is more intonational variability in Canadian-French in comparison to European-French which may assist infants in detecting the boundaries of bisyllabic words at a younger age.

Nazzi *et al*. (2006) further hypothesized that different initial segmentation procedures are involved in languages with different prosodic patterns. The trochaic unit is suggested for trochaic stress-timed languages such as English, while the syllable is suggested to be the unit of segmentation in iambic syllable-timed languages such as French. Thus, studying the prosodic pattern of a language that is different in its prosodic characteristics than those studied thus far may shed light on the expected initial segmentation strategy of infants, as well as provide an opportunity to test the assumption that initial segmentation units differ between languages.

Prosodic structure of Modern Israeli Hebrew

Modern Israeli Hebrew (MH) is an example of a language where, similarly to English, the characteristic stress rhythm is an alternation of strong and weak syllables. Unlike English, the metrical structure of MH is considered to be word-final, weight-insensitive iambic. That is, the assignment of stress does not seem to be affected by heaviness (phonological structure) of syllables and there is no phonemic contrast between full and reduced vowels in strong versus weak syllables (Bolozky, 1982). However, phonetic measurements suggest that a vowel in the stressed syllable may be twice as long as a vowel in an unstressed syllable. Hebrew thus cannot be classified as a syllable-timed language, such as French, in which every syllable is thought to take up roughly the same amount of time (Becker, 2003). Phonological descriptions of stress in MH, based on a limited number of examples chosen from both the nominal and the verbal systems, indicate that stress is assigned to the word-final syllable except in specific environments such as bisyllabic and trisyllabic SEGOLATE nouns,[1] a few environments in the verb system, words influenced by the Yiddish stress patterns (e.g. proper names) and loan nouns and adjectives (e.g. Bolozky, 1982). The position of stress in the nominal system is based on a tripartite distinction between nouns that are assigned ultimate (final) stress in both stem and suffixed form (e.g.

[1] These are penultimately stressed forms in which the vowel *e* predominates – hence their name, *segolates*, after the diacritic mark standing for *e*, *segol*. The two masculine *segolate* patterns yielding bisyllabic nouns are the widespread semantically neutral masculine *CéCeC* (e.g. *kélev* 'dog', *rémez* 'hint'), and the abstract noun pattern *CóCeC* (e.g. *gódel* 'size', *ómek* 'depth'). Feminine *segolates* end with *-éCet*, yielding trisyllabic words. These include both derivational patterns (e.g. *miktéret* 'pipe', *tardémet* 'coma'), as well as feminine inflectional suffixes on present tense participles serving as verbs (*holéxet* 'walking'), nouns (*metapélet* 'therapist') and adjectives (*matmédet* 'continuous'). Finally, some non-*segolate* noun patterns take *segolate* bound construct-state forms, e.g. free *mixlalá* alternates with bound *mixlélet-* 'college'.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

gamád 'dwarf' vs. *gamadím* 'dwarves'), nouns that are assigned penultimate stress in the base form and ultimate stress in the suffixed form (e.g. *náxal* 'river' vs. *nexalím* 'rivers') and those that do not undergo any shifting of stress (e.g. *salát* 'salad' vs. *salátim* 'salads'), which are mostly borrowed words (Bat-El, 1993). Some nouns can also be assigned more than one stress pattern, such as proper names or children's games, with stress location as a salient cue in determining word meaning. As to stress patterns in the verbal system, MH *binyan* (verb pattern) paradigms display ultimate as well as penultimate stress in various environments, but word-final stress is considered most predominant. The most consistent environment for penultimate stress across all *binyan* paradigms is first and second person in past tense – as illustrated in *Qal badákti* 'I checked', *Pi'el tiyálnu* 'we strolled', *Hif'il hivtáxta* 'you + SG promised', and *Hitpa'el hitkaláxtem* 'you + PL showered'. Otherwise, future tense, present tense (non-segolate) and third person past tense verbs tend to take final stress. Some notable alternations are past tense third person feminine singular verbs with regular roots, which take ultimate stress in all *binyanim* (e.g. *Qal sagrá* 'she closed, Tr', *Nif'al nisgerá* 'she closed herself', *Hitpa'el histagrá* 'she closed within herself') except for *Hif'il* (*hisgíra* 'she extradited'); and some stems with irregular roots in future and past tenses, which tend to take penultimate stress – compare, for example, regular *Qal sagrá* 'she closed, Tr' vs. irregular *w*-medial *sáma* 'she put', or regular *Qal yisgerú* 'they will close, Tr' vs. irregular *w*-medial *yavó'u* 'they will come'. The assignment of secondary stress in Hebrew is automatic, that is, every syllable that is one remove from the stressed syllables receives secondary stress (Bat-El, 1993). Thus, some polysyllabic words are considered as carrying secondary stress in addition to the perceptually prominent primary stress. However, judgments regarding secondary stress vary among native speakers, and there are many who fail to identify it (Bat-El, 1993). Overall, while Hebrew appears to be an iambic language, this pattern is not absolute and there are conditions by which it is influenced by lexical class, morphological system or assimilation of other languages. Therefore, in order to study the effect of prosodic patterns on segmentation and language acquisition in Hebrew, we should first estimate the distribution of different stress patterns in the language.

To date, little is known about the general FREQUENCY of different stress patterns in Hebrew. There are no available data regarding the distribution of these patterns in different word classes. Thus, the question whether and how stress pattern distributions influence the development of word segmentation strategies in this language remains open. To the best of our knowledge, only one study provides information on these aspects of Modern Hebrew. Cohen-Gross (1987) tested the morphophonological structure of a total of 16,616 Hebrew nouns and adjectives, taken from a Hebrew dictionary. In this corpus, almost all bisyllabic words (91·5%) were iambs

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

and the rest (8·5%) were trochees. In addition, almost all trisyllabic words were found to have ultimate (final) stress (88%). These findings, although significant, do not provide information about the distribution of the general stress patterns available in everyday speech. Moreover, they do not provide data on prosodic patterns for lexical versus grammatical items or for nouns versus verbs. Furthermore, distributional findings that are based on a dictionary clearly cannot attest as to the differences between the occurrence of various types (the appearance of different words in a corpus of spoken language) and their tokens (the repetition of the same word in the given corpus). This type of research requires mapping predominant stress patterns in general, and in child-directed speech (CDS) in particular. Analyzing CDS is critical since the input directed to infants and toddlers seems to facilitate the process of language acquisition by allowing young language learners to track distributional patterns and regularize them at various levels of linguistic analysis (e.g. Gleitman, Newport & Gleitman, 1984).

Against this background, the present study examines the prosodic characteristics of the Hebrew vocabulary directed to children at the early stages of language acquisition. Our main objective is to examine the distribution of prosodic structure – that is, the number of syllables and predominant stress patterns – in words of different classes that occur in Hebrew CDS. To this end, we present a set of analyses on naturalistic corpora as described below – focusing on the number of syllables, assignment of stress, and categorization into word classes. Specifically, the first analysis aims to examine the distribution in types and tokens of words of different lengths, in order to determine the predominant category of word length in Hebrew CDS. The analysis will allow us to extend our knowledge about the actual frequency of different word lengths in Hebrew, and to assess to what degree CDS correlates with or differs from other Hebrew databases. A second analysis considers the distribution of all tokens and types into different word classes as well as the internal distributions of word length in each of these classes. This mapping lays the basis for our main set of analyses that consider the occurrence of different stress patterns within bi- and polysyllabic words and in different lexical categories in order to characterize the predominant stress patterns of adult speech in various situations of parent–child interactions. These will allow us to assess whether the predominant prosodic structure is equally observed in all lexical categories and subclasses.

METHOD

The corpora

The database for this study consists of two corpora of naturalistic samples that were accessed from the CHILDES database (MacWhinney, 2008). All

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

speech samples which included child-directed speech were transcribed using standard CHAT procedures as specially adapted for Hebrew.

Data were collected from middle- to upper-middle-class adult caregivers of Hebrew-speaking children between the ages of 0;9 to 3;0 during daily interactions – for example, conversations about everyday events, playing, painting, looking at books, picture book reading and story-telling and retelling. These samples are based on the Berman Longitudinal corpus – which includes four longitudinal datasets (a total of 392 sessions) that were collected as part of a cross-linguistic project headed by Professor Ruth Berman of Tel-Aviv University (1988–1991),[2] and on the Ravid corpus – a longitudinal sample (19 sessions) collected by Professor Dorit Ravid, Tel-Aviv University, that includes the speech directed to two of her children (1980–1985).

In general, recordings took place at each child's home environment, when the children were playing with their caregivers in their bedroom or the living room. The adults who are recorded in these samples were the five pairs of parents (mainly the mothers) and in one case an additional investigator, the aunt of the recorded child. Other family members (grandmother, grandfather, uncle) occasionally participated in the sessions, as well as neighbors and friends of the family who sometimes played with the children. That is, speech samples were collected from the same group of people for each child across the sessions. Most adult speakers were native monolingual Hebrew speakers (with the exception of some of the grandparents). Thus, the longitudinal corpus provides extensive information on recurring interactions in specific families. In terms of the variation of adult speech, the entire corpus includes samples from around twenty-three different participants, but the bulk of CDS utterances (61%) were those of the five mothers.

Analytic procedures

A list containing the surface forms of the database was extracted, using the FREQ command (MacWhinney, 2008). The list of words was then imported into an Excel spreadsheet, to allow for multiple levels of coding for each type – number of syllables, stress pattern and word classes. These analyses were then applied to the entire database in order to allow for statistical comparisons.

[2] Funding for data-collection and transcription of these materials was provided by grants to Ruth Berman, Tel Aviv University, and Jürgen Weissenborn, Max-Planck Institute for Psycholinguistics, Nijmegen, from the German–Israel Binational Science Foundation (GIF) – 1988 to 1991 – and from the Deutsche Forschungsgemeinschaft (DFG) – 1988 to 1990 – for the crosslinguistic study of early language acquisition in French, German and Hebrew.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

The database was analyzed by one of the authors – a graduate student in speech and language pathology, with reliability checks done by a second author, a graduate student in linguistics. Both are native speakers of Hebrew and experienced coders. Inter-coder agreement was high, and differences were resolved after discussion. Since there are no available computerized systems for analyzing continuous speech in Hebrew, all analyses – including those related to stress patterns – were carried out based on native judgments.

Number of syllables. Coding of word length in terms of number of syllables included three explicit levels: (1) monosyllabic words, for example *sof* 'end', *ec* 'tree', *dod* 'uncle', *pe* 'mouth'; (2) bisyllabic words, for example, *bubá* 'doll', *ába* 'dad', *raxók* 'far', *ayéf* 'tired'; and (3) trisyllabic words, for example, *zaxárti* 'remembered + 1st + SG, *kacéfet* 'cream', *makólet* 'supermarket'. An additional level, of polysyllabic words, included words of four syllables and more, such as *televízia* 'television', *sufganiyá* 'doughnut', *mištaamemím* 'getting-bored + PL'.

Word classes. Berman (2001) suggests that although the distinction between closed- and open-class elements of vocabulary is intuitively appealing, in many ways it constitutes too sharp a division. Berman proposes that word-class categorization is best characterized as a cline or continuum and posits a division between three major classes, as follows: (1) closed class (CC); (2) between class (BC); and (3) open class (OC). Thus, closed-class items include paradigms such as articles, demonstratives and personal pronouns such as *ha* 'the', *ze* 'this', and *ani* 'I', as well as more open-ended lexical classes such as prepositions, e.g. *be* 'in', *me* 'from', *le* 'to', coordinating and subordinating conjunctions, such as *ve, aval, o, še* (the equivalents of English *and, but, or, that*), negators (*lo* 'no'), quantifiers (numbers, *kol* 'all', *kama* 'some', *harbé* 'many'), question words (*mi* 'who', *ma* 'what', *matáy* 'when'), etc. The open-class category includes nouns, verbs and adjectives, such as *basár* 'meat', *baxú* 'cried' and *cahóv* 'yellow', and a fourth category of all and only adverbs derived by the instrumental prepositional *be-* plus an abstract stative nominal, such as *bi-mehirút* 'with-quickness', *bi-zehirút* 'with carefulness' (the equivalents of English derived *-ly* adverbs *quickly, carefully*).

The innovative feature of Berman's (2001) proposal is positing an additional group of elements termed 'between-class items' to capture the sense that these lie somewhere in between fully grammatical and fully contentive lexical elements. These are words like adverbs and the various elements clumped together under the general term of 'discourse markers', such as *gam* 'also', *rak* 'just', *afílu* 'even', *tov* 'okay', *yófi* 'great', *malé* 'a lot of'. Such items typically do not have an autonomous semantic interpretation – like content words – and their meaning is relative not only to the syntactic environment in which they occur – like functional items – but also

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

to the particular TEXTUAL OR DISCURSIVE CONTEXT in which they occur. They constitute a kind of 'elsewhere' class, and often apply to set expressions of more than a single orthographic word, lumped together in traditional grammars under the heading of 'adverbs' or 'adverbials'. In this sense, it should be hard to predict their prosodic structure since they combine features from the other two groups. Since many of the items in this category can only be identified within the entire utterance, the COMBO command (MacWhinney, 2008) was applied in order to retrieve the exact context. It should be noted that another group of items that could not be classified into either one of the above lexical classes was also analyzed and labeled as 'Others'. It consisted of vocatives (*hópa* uttered after an incident of falling, *kúku* uttered during a game of hide-and-seek, *brrrr* indicating the sound of a car), animal voices (*gagaga* for a duck, *haw* for a dog), fillers (*ahm*, *ah*), and other pragmatic and emotional expressions.

Stress patterns. For the final set of analyses, coding included specifying word stress patterns for bisyllabic words and the location of stress in relation to the final syllable for tri- and polysyllabic words – in all four categories.[3] Bisyllabic words were tagged either as carrying trochaic (strong–weak) stress – for example, *ába* 'dad', *náxa* 'rested + FEM + 3rd' – or iambic (weak–strong) stress, for example *bubá* 'doll', *axál* 'ate + MASC + 3rd', *raxók* 'far'. Tri- and polysyllabic words were divided into three groups according to location of primary stress – either on the final syllable (labeled as [FIN]), as in *sufganiyá* 'doughnut' and *mištaamemím* 'getting-bored + PL', on the penultimate syllable (labeled as [PENUL]), as in *zaxárti* 'remembered + 1st', and on the antepenultimate syllable (labeled as [ANTE]), as in *tráktorim* 'tractors'.

RESULTS

The adult input in the Hebrew longitudinal corpus yields a total of 228,946 tokens (words) and 8,075 types (different word forms) over 57,536 utterances (MLU in words = 3·9). Only around 15% of all utterances were one word in length. All word tokens and types were analyzed according to number of syllables, lexical class, and stress pattern, based on the explicit categories of analysis noted above. Note that we consider a very specific definition of 'word type' – based on surface form. Words with a different morphological structure, as for example, *zoxer* 'remember + SG' and *zaxar* 'remembered + SG', are counted as two different word types.

[3] There is general agreement that the assignment of stress to monosyllabic words is dependent on their position in continuous speech; their perceptual dominance can only be determined in relation to their lexical category (either open- or closed-class items) combined with the stress patterns and lexical categorizations of surrounding syllables. For these reasons, the current study excludes monosyllabic words from stress pattern analyses.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

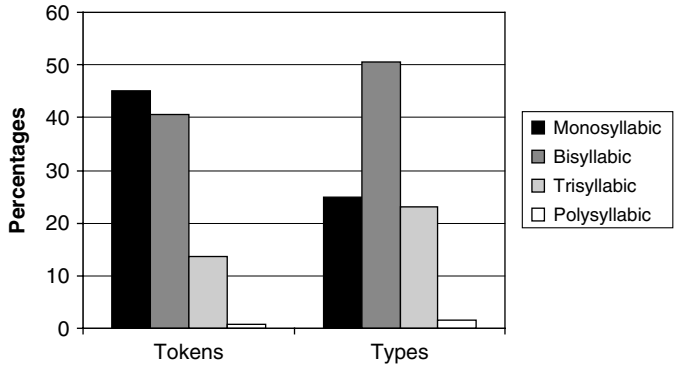


Fig. 1. Percentages of word length categories in Hebrew corpora: Tokens vs. types.

Distribution of word length categories

Our first analysis aims to determine the predominant category of word length in Hebrew CDS. The distributions in types and tokens of words of different lengths in our corpus are presented in Figure 1.

Figure 1 shows that, across the database, there are clearly more mono- and bisyllabic word TOKENS ($M = 44{\cdot}8$ and $M = 40{\cdot}7$ respectively) as compared to trisyllabic word tokens ($M = 13{\cdot}6$). In comparison, bisyllabic word TYPES were found to be most frequent ($M = 50{\cdot}5$) in CDS, while mono- and trisyllabic word types each take up around one quarter of the data ($M = 24{\cdot}6$ and $M = 23{\cdot}2$ respectively). Kruskal–Wallis analysis of variance tests showed differences to be statistically significant for both word tokens ($\chi^2 = 603{\cdot}8$, $p < 0{\cdot}01$) and word types ($\chi^2 = 560{\cdot}6$, $p < 0{\cdot}01$). Additional Mann–Whitney tests revealed that the means for the three variables (mono-, bi- and trisyllabic words) were significantly differentiated ($p < 0{\cdot}05$ for all pairs). These analyses show that young Hebrew-acquiring children are exposed mainly to monosyllabic word tokens and to bisyllabic word types. In other words, although monosyllabic words are frequent in terms of their overall number in the corpus, they are confined to a relatively small set of word forms that tend to repeat themselves. In contrast, bisyllabic words are more diverse, and polysyllabic words are rare in Hebrew CDS, both in type and in token frequency. These distributions of mono-, bi- and trisyllabic types versus tokens may be attributed to their classification into different lexical categories, as explained below.

Distribution of lexical categories

We calculated the relative distribution of the three categories used for analysis (open-class, closed-class, between-class) in the database, by token

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

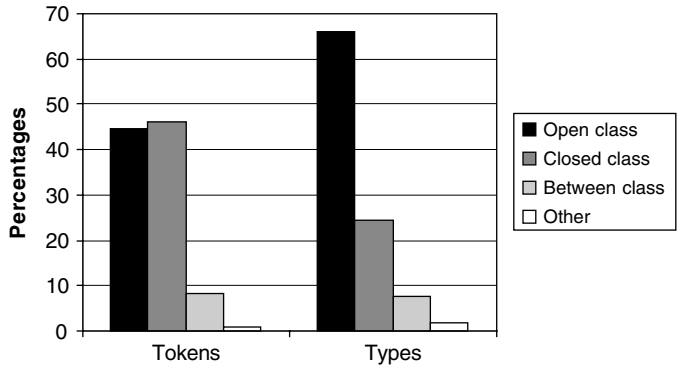


Fig. 2. Percentages of different lexical classes in Hebrew corpora: Tokens vs. types.

and type, as presented in Figure 2. The distributions in Figure 2 show that closed-class items are frequent in tokens (around 46%), but restricted in types – less than 25%. In contrast, while open-class items take up as much as 45% of all word tokens, they constitute 66% of all word types. Between-class items constitute less than 8% of all word tokens and of all word types, and the items included under the heading 'Other' take up only between 1 and 2% of all data. In other words, the open-class category is the most varied of the four groups, with the highest ratio between types and tokens. Kruskal–Wallis analysis of variance tests showed these differences to be statistically significant for both word tokens ($\chi^2 = 940{\cdot}8$, $p < 0{\cdot}01$) and word types ($\chi^2 = 1040{\cdot}7$, $p < 0{\cdot}01$). Again, a set of Mann–Whitney tests revealed that the means for the four categories (open-class, closed-class, between-class and other) were significantly differentiated ($p < 0{\cdot}01$ for all pairs).

Distribution of lexical categories according to word length

The next set of analyses aims to determine the predominant category of word length in each of the lexical categories specified above. Figures 3a and 3b compare percentages for each category in terms of tokens and types, showing that open-class items are mainly bi- and trisyllabic. Kruskal–Wallis analysis of variance tests revealed the differences between the three word-length categories to be statistically significant in both tokens ($\chi^2 = 727{\cdot}8$, $p < 0{\cdot}01$) and types ($\chi^2 = 741{\cdot}8$, $p < 0{\cdot}01$). Closed-class items are mainly monosyllabic, again with a Kruskal–Wallis analysis of variance showing the differences between the three word-length categories to be statistically significant for tokens ($\chi^2 = 733{\cdot}7$, $p < 0{\cdot}01$) and types ($\chi^2 = 712{\cdot}2$, $p < 0{\cdot}01$). Note that there were more bisyllabic types in the closed-class category as compared to tokens ($M = 38$ and $M = 20{\cdot}8$, respectively).

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

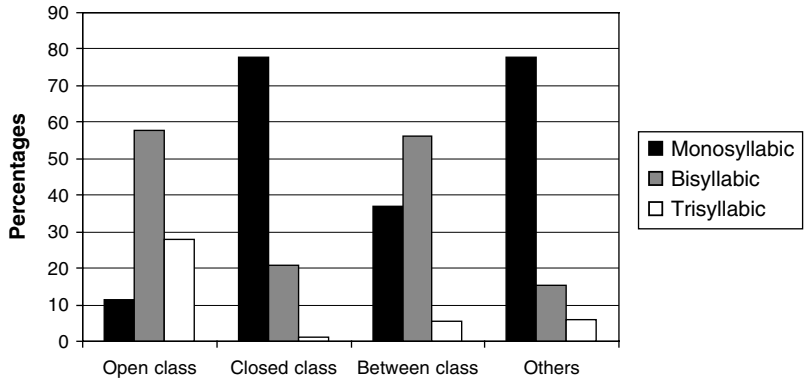


Fig. 3a. Percentages of mono-, bi- and trisyllabic words by lexical category (tokens).

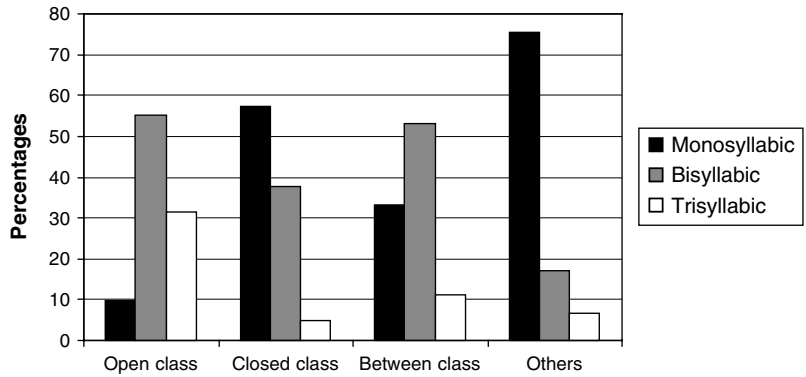


Fig. 3b. Percentages of mono-, bi- and trisyllabic words by lexical category (types).

Between-class items are mainly mono- and bisyllabic with the same highly significant differences for both tokens ($\chi^2 = 633{\cdot}3$, $p < 0{\cdot}01$) and types ($\chi^2 = 633{\cdot}8$, $p < 0{\cdot}01$). Finally, the bulk of items tagged as others is monosyllabic.

As these figures show, bisyllabic word types are highly predominant in the three major lexical classes and especially in the open-class category (58% and 55% for both token and type, respectively). However, monosyllabic items are mainly predominant in the closed-class group for both token (78%) and type (57%) distributions. Also, the characterization of the category of 'others' remains constant as predominantly monosyllabic in terms of both tokens ($M = 78$) and types ($M = 75{\cdot}4$). These results throw light on the general findings presented above with respect to the interaction between word length on the one hand and word types and tokens on the other: most

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

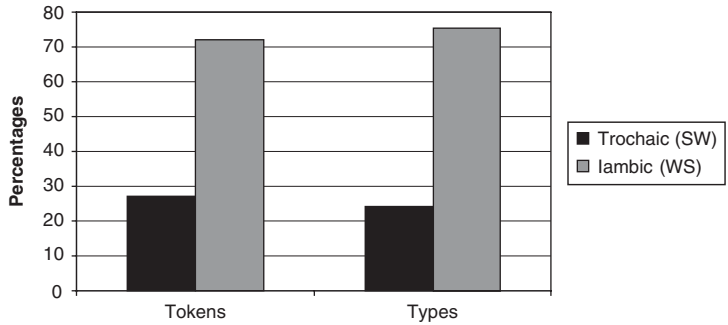


Fig. 4. Percentages of trochaic (SW) versus iambic (WS) stress patterns in bisyllabic words: Tokens versus types.

of the bisyllabic words belong to the open-class category, a class of items that is traditionally considered as containing numerous members with relatively low frequency (that is, high type–token ratio) as distinguished from the small set of closed-class items that are of high frequency (Berman, 2001).

Distribution of stress patterns in bisyllabic words

We now turn to the analyses at the heart of this study – the distribution of stress patterns in bisyllabic words. These analyses aim to characterize predominant stress patterns in Hebrew CDS. Figure 4 presents the distribution of trochaic (SW) versus iambic (WS) stress patterns in bisyllable words. Strikingly, iambic (WS) stress emerged as the most predominant pattern in word tokens (72·2%) and types (75·5%).

A Kruskal–Wallis analysis of variance showed the differences between the distributions of the two stress patterns to be statistically significant for both word tokens ($\chi^2 = 416{\cdot}5$, $p < 0{\cdot}01$) and word types ($\chi^2 = 419{\cdot}3$, $p < 0{\cdot}01$). In order to ensure that the iambic stress pattern is predominant in the input directed even to the youngest age group, we divided our database into three subgroups, each consisting of a relatively similar number of sessions and representing relatively parallel age ranges: Group I includes input to children between 0;9 and 2;0; Group II includes input to children between 2;0 and 2;5; and Group III includes input to children from 2;5 to 3;0. The division yields three datasets of highly equivalent sample sizes in terms of number of sessions (Group I = 104 sessions, Group II = 102 sessions, Group III = 74 sessions) and relatively similar distributions in terms of word tokens (Group I = 61,942, Group II = 93,952, Group III = 73,052).

The results indicate an increase in the use of iambic words in the input as a function of age, as shown by the data presented in Table 1. The figures in

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

TABLE 1. *Token and type distributions (in percentages) of stress patterns in bisyllabic words in longitudinal Hebrew corpora*

	TROCHAIC (SW)		IAMBIC (WS)	
	Tokens	Types	Tokens	Types
Group I Age: 0;9–2;0	28·3	25·4	69·5	72·1
Group II Age: 2;0–2;5	27·6	23·9	72·9	76·9
Group III Age: 2;5–3;0	24·8	22·6	75·1	77·9

Table 1 clearly show that iambic stress is the predominant pattern in the input directed to all three age groups. However, the input to the youngest children contains a greater proportion of trochaic word TOKENS compared to the input directed to the oldest children ($\chi^2 = 6{\cdot}58$, $p < 0{\cdot}05$) and significantly more trochaic word TYPES compared to both children between 2;0 and 2;5 ($\chi^2 = 6{\cdot}4$, $p < 0{\cdot}05$) as well as in comparison to the input addressed to the oldest age group ($\chi^2 = 9{\cdot}9$, $p < 0{\cdot}01$).

Moreover, a significant difference emerged between groups I and II in both iambic word TOKENS ($\chi^2 = 21{\cdot}7$, $p < 0{\cdot}01$) and word TYPES ($\chi^2 = 35{\cdot}4$, $p < 0{\cdot}01$) as well as between groups I and III in iambic word TOKENS ($\chi^2 = 10{\cdot}3$, $p < 0{\cdot}05$) and word TYPES ($\chi^2 = 33{\cdot}3$, $p < 0{\cdot}01$). A difference also emerged between group II and III, but only in iambic word TOKENS ($\chi^2 = 7{\cdot}6$, $p < 0{\cdot}05$).

The next set of analyses aims to determine whether these distributions are equally observed in all lexical categories. Data are presented in Figures 5a and 5b. Strikingly, it seems that items belonging to the group tagged as 'other', which consists of non-lexical types such as *kúku* 'peek-a-boo' and *hópa* 'whoopsidaisies', are predominantly trochaic (around 90% in both tokens and types). As for the three major lexical classes, the distributions in Figures 5a and 5b indicate that the common stress pattern for bisyllabic words is iambic (WS) – between 70 and 80% in tokens and between 60 and 80% in types. These differences are statistically significant for both open-class tokens ($\chi^2 = 416{\cdot}03$, $p < 0{\cdot}01$) and types ($\chi^2 = 417{\cdot}7$, $p < 0{\cdot}01$), for closed-class tokens ($\chi^2 = 340{\cdot}8$, $p < 0{\cdot}01$) and types ($\chi^2 = 410{\cdot}8$, $p < 0{\cdot}01$) and also for between-class tokens ($\chi^2 = 362{\cdot}1$, $p < 0{\cdot}01$) and types ($\chi^2 = 207{\cdot}6$, $p < 0{\cdot}01$). Thus, a very consistent pattern emerges for all three lexical categories. However, significant differences were also found when comparing the relative proportion of iambic stress between the three major lexical classes in both tokens ($\chi^2 = 58{\cdot}05$, $p < 0{\cdot}01$) and types ($\chi^2 = 185{\cdot}7$, $p < 0{\cdot}01$). A series of Mann–Whitney tests revealed that there are significantly MORE

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

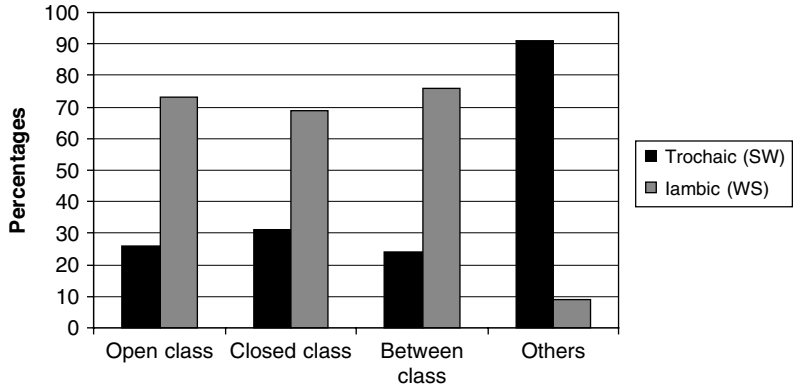


Fig. 5a. Percentages of trochaic (SW) versus iambic (WS) stress patterns by lexical category (tokens).

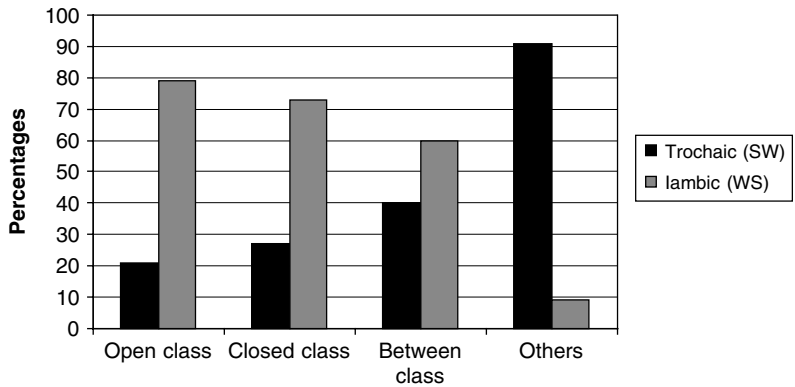


Fig. 5b. Percentages of trochaic (SW) versus iambic (WS) stress patterns by lexical category (types).

iambic word tokens ($Z = -2{\cdot}7$, $p < 0{\cdot}01$) and word types ($Z = -7{\cdot}8$, $p < 0{\cdot}01$) in the open-class compared with the closed-class category; there are significantly LESS iambic word TOKENS ($Z = -5{\cdot}9$, $p < 0{\cdot}01$) in the open-class compared with the between-class category but significantly MORE open-class iambic TYPES ($Z = -12{\cdot}6$, $p < 0{\cdot}01$); there are significantly LESS iambic word TOKENS in the closed-class compared with the between-class category ($Z = -6{\cdot}8$, $p < 0{\cdot}01$) and significantly MORE closed-class iambic TYPES ($Z = -7{\cdot}8$, $p < 0{\cdot}01$). These results suggest that the iambic stress pattern is most frequent and most diverse in the open-class category.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

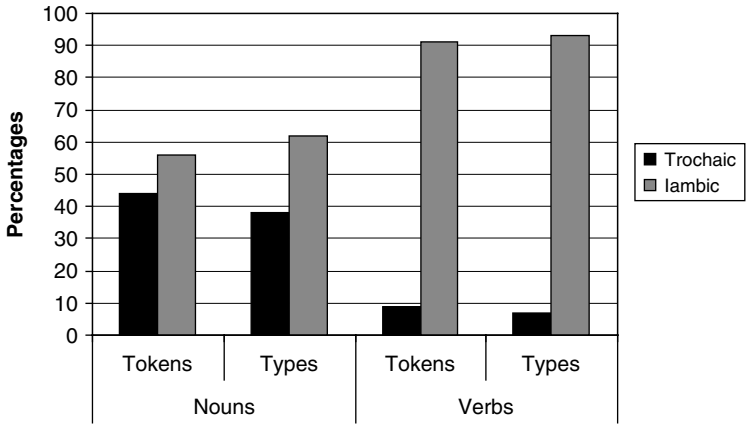


Fig. 6. Percentages of trochaic (SW) and iambic (WS) stress patterns in nouns and verbs: Tokens versus types.

As noted above, studies on English show that there is some variation in the distribution of stress patterns between different subcategories of the open class. In our sample, the two major subclasses (constituting around 80% of all tokens and types) are nouns and verbs. And within each of these subclasses, above 50% of all tokens and types are bisyllabic. Our final analysis compares the distributions of stress patterns in these two categories.

As shown by Figure 6, the iambic stress pattern is predominant for both nouns and verbs. This was found to be statistically significant for noun tokens ($\chi^2 = 71{\cdot}8$, $p < 0{\cdot}01$) and types ($\chi^2 = 267{\cdot}8$, $p < 0{\cdot}01$) as well as for verb tokens ($\chi^2 = 417{\cdot}7$, $p < 0{\cdot}01$) and types ($\chi^2 = 417{\cdot}8$, $p < 0{\cdot}01$). However, there are significantly more iambic verbs than nouns in both tokens ($\chi^2 = 365{\cdot}3$, $p < 0{\cdot}01$) and types ($\chi^2 = 377{\cdot}9$, $p < 0{\cdot}01$).

Distribution of stress patterns in trisyllabic words

Additional information on the distribution of stress patterns in Hebrew can be extracted from the analysis of trisyllabic words. Thus, although this group amounts to only around 13% of all words in TOKEN frequency, it makes up almost 25% of all word TYPES. Stress patterns in trisyllabic words were analyzed by the position of primary stress – either on the first (ante-penultimate), penultimate or final (ultimate) syllable. As shown in Figure 7, trisyllabic words divide up relatively equally in terms of tokens of words that carry primary stress on the penultimate syllable and those that carry it on the final syllable. Moreover, trisyllabic words that carry stress on the

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

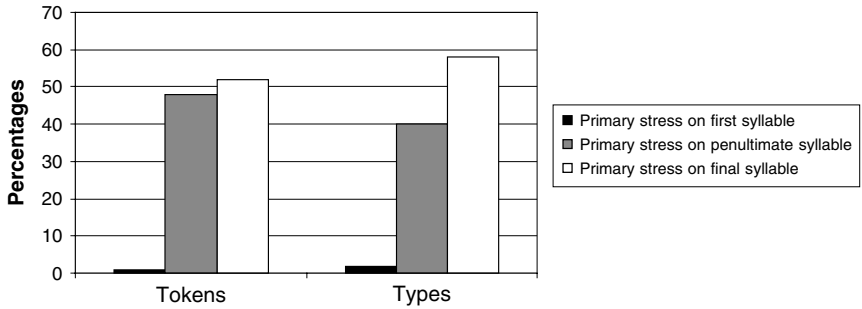


Fig. 7. Percentages of trisyllabic words with primary stress on the first, penultimate and final syllable: Tokens versus types.

final syllable constitute almost 60% ($M = 58\%$) of all types. The proportion of words that carry stress on the first syllable is negligible, reaching as little as 1% of all word tokens and only 2% of all word types.

A Kruskal–Wallis analysis of variance comparing the words that carry primary stress on the penultimate syllable and words that carry primary stress on the final syllable yielded a significant difference in both word tokens ($\chi^2 = 29{\cdot}3$, $p < 0{\cdot}01$) and word types ($\chi^2 = 280{\cdot}2$, $p < 0{\cdot}01$). Taken together, the numbers presented in Figures 4 and 7 indicate that, across the board, stress in Hebrew CDS is assigned to the final syllable (in both tokens and types).

DISCUSSION

The main goal of the study was to provide data about the distribution of prosodic patterns in Hebrew. Recent work has brought forth new findings, suggesting that infants and toddlers use distributional learning of the common prosodic patterns in their language as a tool for word segmentation and for assigning words to different lexical or grammatical categories (e.g. Jusczyk, Houston & Newsome, 1999; Nazzi *et al.*, 2006; Shi *et al.*, 1999). Given that different languages expose their novice learners to different prosodic patterns (e.g. trochaic vs. iambic languages) and to different distributions of prosodic cues, the first step towards studying how language-acquiring infants use the prosodic patterns of their native language – in the present case, Hebrew – requires a detailed description of the characteristic prosodic input.

Against this background, we addressed the issue of the frequency of two prosodic cues – word length and stress pattern – in Hebrew child-directed speech. We also inquired as to whether the distribution of these prosodic

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

cues in early input differs across lexical categories and subcategories. For this purpose we conducted an analysis of the interaction of prosodic patterns and three lexical classes (open class, closed class and between class) as well as with two OC subclasses: nouns and verbs.

Our findings suggest that: (a) young Hebrew-acquiring children are mainly exposed to a limited but frequent set of monosyllabic words and to a diverse group of bisyllabic words; (b) word-length distributions differ across lexical classes (OC, CC and BC); and (c) the common stress pattern in Hebrew CDS is word-final. In the remainder of the discussion, we elaborate and clarify the implications of these findings in light of issues in early language development.

The distribution of prosodic patterns in Hebrew CDS

Number of syllables. Our first finding is that mono- and bisyllabic words are the most frequent in Hebrew CDS. In terms of word TOKENS, monosyllabic words are of the highest frequency (constituting 45% of all tokens) followed by bisyllabic words (41% of all word tokens), with trisyllabic words comprising only about 14% of all word tokens. Strikingly similar frequencies were found in an independent study in which number of syllables was analyzed in a corpus consisting of 320 narrative and expository spoken texts produced by school children, adolescents and adults (Nir-Sagiv, 2005). Across this database, monosyllabic words were found to constitute as high as 45% of all tokens in narratives and 46% in expository texts, and bisyllabic words made up 34% of all tokens in narratives and 31% in expository texts. The majority of the remaining tokens were trisyllabic, at 18% of all tokens in narratives and 19% of all tokens in expository texts. The high correspondence between the findings of the two studies suggests that the make-up of Hebrew in general consists of mono- and bisyllabic words, and that Hebrew CDS is not as differentiated from adult speech in this respect – although it does seem that trisyllabic words are to some extent more frequent in the context of monologic texts produced by schoolchildren, adolescents and adults.

In terms of word TYPES, results show that bisyllabic words are more diverse (51% of all types consist of bisyllabic words) compared with mono- and trisyllabic word types (25% and 23% respectively). Previous studies report on similar trends for both Hebrew and English: Cohen-Gross' (1987) descriptive analysis of the Hebrew dictionary (nouns and adjectives) found that monosyllabic word types were limited in frequency – about 2% (only 192 words) – with the rest consisting of mainly bi- and trisyllabic words (50%). For English, Cutler & Carter (1987) suggest that although monosyllabic words are frequent in spoken English in tokens, the most common type is bisyllabic.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

Our findings may be used to explicate specific issues in language acquisition in terms of, for example, early speech production. Thus, in the first extensive investigation of the acquisition of phonological knowledge in Hebrew, Ben-David (2001) found that mono- and bisyllabic words appear earlier in children's productive lexicons, suggesting that these unmarked structures are acquired earlier in the course of development. Our findings, however, provide a language-specific distributional perspective, pointing to features of child-directed input as a possible substantive factor influencing children's early productions.

Word length in different lexical classes. An additional goal of our study was to inquire whether prosodic information in CDS could support a distinction between different word classes. Our findings suggest that functional closed-class items as well as the non-lexical productions termed as 'others' consist mainly of monosyllabic words (almost 80% and 60% of all tokens and types in the closed-class and almost 80% of the 'others' category); while semantically complex open-class items consist mainly of bi- and trisyllabic words (about 87% of all tokens and types). As to items belonging to the between-class category, our results show a majority of bisyllabic word tokens (56%) and types (53%) as well as highly frequent monosyllabic tokens and types (37% and 33% respectively). Thus, while both open- and closed-class items are more distinctly differentiated in word length, between-class items show a more mixed tendency. In other words, it seems that the semantic ambiguity of the members of this group is also reflected in its prosodic characteristics. Thus, our results provide new support to the claim that lexical categories are frequently marked by different distributions of word length. Moreover, our results are consistent with previous findings for English, where Cutler & Carter (1987) reported that over 86% of the closed-class items in their corpus were monosyllabic, and most (above 50%) of the open-class items were polysyllabic. Furthermore, findings in English and Turkish CDS confirm that lexical items are more likely to be polysyllabic (Shi *et al.*, 1998).

Taken together, the findings for the three word classes suggest that, in Hebrew CDS, number of syllables can serve as a perceptual cue for categorizing words. From the data presenting number of words of differing length (Figure 1) and out of different lexical categories (Figure 3a), we can calculate statistical probabilities for the interaction of word length and word class as presented in Table 2.

The probability that, for example, a monosyllabic word occurring in CDS would belong to the closed-class category is the proportion of monosyllabic closed-class tokens out of all monosyllabic tokens. The percentages in Table 2 show that the probability for a monosyllabic word heard by a Hebrew-acquiring child to belong to the closed-class category is around 79·5%, the chance for bisyllabic words in CDS to belong to the open-class

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

TABLE 2. *Statistical probabilities (in percentages) for assignment of word class as a function of word length*

	Monosyllabic words	Bisyllabic words	Trisyllabic words
Closed class	79·5	23·7	4·5
Between class	6·9	11·5	3·3
Open class	11·3	63·7	91

category is over 60%, and the probability that a child identifies open-class items when hearing trisyllabic words is 91%. Thus, these data support the hypothesis that once the child learns to associate word length with word class, this acoustic perceptual cue may guide her in assigning a specific category to newly heard words.

Stress patterns. To the best of our knowledge, this is the first extensive corpus-based analysis that confirms previous claims about the predominance of word-final stress patterns in Hebrew (Bat-El, 1993; Bolozky, 1982; Cohen-Gross, 1987). Our results indicate that Hebrew CDS is characterized mainly by word-final stress: most of the bisyllabic words in the database are iambic (72·2% and 75·5% of all word tokens and types, respectively). Moreover, trisyllabic words begin mostly either with weak syllables (primary stress on the penultimate syllable) or with initial secondary stress (primary stress on the final syllable), and rarely begin with initial primary stress (only 1% of all word tokens).

It is possible to assume that such biased frequencies and distributions have an effect on the acquisition of stress in terms of speech perception and production. For example, we may predict that a Hebrew-speaking child would produce mainly iambic strings from the very beginning of the bisyllabic stage. Accordingly, findings for Hebrew-acquiring children suggest that final (and thus, in most cases, stressed) syllables are the first to be produced in the earliest stages of acquisition (Ben-David, 2001). However, such predictions are attenuated by the significant change in the distribution of stress patterns in CDS as a function of age. Thus, our developmental analysis revealed that although iambic words are clearly predominant across the database, they are more frequent in adult speech directed to older children (2;5–3;0) compared to the two younger age groups.

In terms of the distribution of stress patterns across different lexical classes, our results show that the iambic pattern is common in all three classes analyzed here. However, it is most predominant in the open-class category in word TYPES (79% of all types as compared to 73% and 60% in the other two major classes). As noted earlier, open-class types in themselves are most predominant (66% of all words). In other words, iambic, bisyllabic open-class types are highly frequent in our database of Hebrew CDS. Type

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

frequency is viewed as a key factor in a usage-based approach to language acquisition (e.g. Tomasello, 2003). Thus, although the iambic stress pattern is predominant across the board, it may still serve as a cue to young Hebrew-acquiring children for distinguishing between different lexical categories.

An interesting finding of the present study is that the limited set of non-lexical items termed 'others' shows a significantly different pattern of results. Most bisyllabic productions in this group (around 90%) are trochaic. To the best of our knowledge, no previous research reported on the stress patterns of this group of communicative vocal productions. The function of such non-lexical (mainly isolated) items consisting of a first stressed syllable and a last weak syllable seems to be to emphasize semantic–pragmatic aspects in a play situation, as well as to call for infants' attention during interaction (e.g. '*hópa*' uttered after an incident of falling, *kúku* uttered during a game of hide-and-seek, *zúmzum* indicating the sound of a fly, *yála* indicating end of one interaction and the beginning of another, and diminutive forms such as *mámuš*, *mámi*, *zísi*, *búbi* indicating affection). It has been suggested that young children acquiring trochaic languages may be predisposed to extracting and storing stressed and final syllables which are perceptually more salient (due to changes in loudness, fundamental frequency and duration, in the stressed syllable and lengthening of the final syllable) – and thus to treat those syllables as their initial representation for words while omitting weak, non-final syllables (Echols & Newport, 1992). Turning back to Hebrew CDS, the use of frequent vocal trochaic templates may have a perceptual advantage in comparison to iambic templates, because it consists of a stressed syllable followed by a weak syllable, and thus both may be acoustically salient. This perceptual advantage may be communicatively used by adult speakers for arousing or maintaining attention in infants and young toddlers, even though the general pattern of words in the Hebrew language is iambic. However, it is important to note that the use of such vocal productions was observed mainly at an early age, in CDS directed to children at the age of 0;9 to 2;0, and far less in the older age ranges, suggesting a developmental change in which less non-lexical productions and more lexical items appear in the input directed to toddlers at the age of 2;0 to 3;0. This coincides with the significant DECREASE in trochaic words in CDS observed between the three age ranges.

As for the sublexical distinction between nouns and verbs, the iambic stress pattern was found in the present study to be more frequent in verbs (91%, 92·7% of all tokens and types, respectively), although it remained predominant also for nouns (56%, 61·6% of all tokens and types, respectively). These findings differ from the findings for English, where the common stress pattern for nouns is trochaic, while most verbs are iambic (Kelly & Bock, 1988).

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

Implications and further research

Previous studies on segmentation have shown that proficient adult speakers of trochaic languages (e.g. English) as well as infants use stress as a perceptual cue for identifying (content) word beginnings, positing word onset at each strong syllable (Cutler & Norris, 1988; Jusczyk, Houston & Newsome, 1999). It is not clear, however, whether the development of such a strategy in a trochaic language is biased by the acoustic saliency of the strong–weak stress pattern, or whether it is mainly dependent on exposure to the specific strong–weak type of input (e.g. Echols & Newport, 1992; Echols *et al*., 1997). Thus, considering the question of how stress patterns guide infants in word segmentation in iambic languages is important for our understanding of the role of language-specific versus acoustic–universal procedures. Such studies are rare and their results are somewhat conflicting. Polka & Sundara (2003) found that Canadian French-learning infants aged 0;7·15 segmented bisyllabic iambic words from fluent speech in French, but not trochaic words in English. In contrast, Nazzi *et al*. (2006) found that infants aged 1;0 did not segment iambic words from fluent speech, but were able to segment separate syllables of these words whether they were weak or strong. These results were in agreement with preceding studies in French adults suggesting that the syllable is the unit for speech segmentation and processing (e.g. Peretz, Lussier & Béland, 1998). Thus, Nazzi *et al*. (2006) raised the possibility that the segmentation of bisyllabic words in French begins with the unit of the syllable. According to this suggestion, languages differing in their prosodic characteristics and rhythmical classes as English and French impose initial dissimilar segmentation units on infants, when the trochaic unit (i.e. sequences of syllables that begin with a strong syllable) is expected for English and the syllable unit is suggested for French.

As for Hebrew, our findings indicate that infants and young children are frequently exposed to an iambic (weak–strong) rhythmic pattern in bisyllabic words, and mainly to ultimate stress in trisyllabic words. It is plausible to assume that Hebrew-acquiring infants are attuned to these common stress patterns, and in combination with other acoustic cues – such as phonetic and phonotactic features as well as syllable transitional probabilities – they rely on them for word segmentation. Thus, relying on the stressed syllable as a cue for word beginning might not be as efficient a tool for the segmentation of words by young Hebrew learners as it is for young English learners. Alternatively, Hebrew-learning infants may initially use other procedures such as parsing each syllable separately, using the stressed syllable as a cue for word ending or detecting the entire common rhythmic pattern (e.g. the common weak–strong pattern of bisyllable words) in the speech stream. Further research, however, is needed in order to address this issue. Also, our findings for the different distributions of stress patterns

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

when comparing nouns and verbs raise interesting questions concerning initial segmentation strategies for these two subclasses in Hebrew. Finally, the findings of the current study will serve as the basis for interpreting the results of ongoing research that aims to assess whether Hebrew-acquiring young infants develop a preference for the predominant stress pattern in their language or for the acoustically salient strong–weak stress pattern (Segal, in progress). On the one hand, the trochaic pattern is less frequent in Hebrew and thus should not be preferred based on language-specific exposure. On the other hand, the trochaic items consisting of a stressed syllable at its onset and a final weak syllable can be perceptually more salient, and thus may universally attract infants' attention (Echols & Newport, 1992). One way to assess these two possibilities is to study a language with an iambic stress pattern (Echols *et al.*, 1997). Thus, coupled with the distributional data of prosodic patterns in Hebrew CDS, the results of infants' stress pattern preference may shed new light on the interaction of language specific features and reliance on general auditory mechanisms in the earliest stages of language acquisition.

REFERENCES

Arciuli, J. & Cupples, L. (2004). Effects of stress typicality during spoken word recognition by native and nonnative speakers of English: evidence from onset gating. *Memory and Cognition* **32**, 21–30.

Bat-El, O. (1993). Parasitic metrification in the Modern Hebrew stress system. *The Linguistic Review* **10**, 189–210.

Becker, M. (2003). Hebrew stress: Can't you hear those trochees? In E. Kaiser & S. Arunachalm (eds), *Proceedings of the PLC* **26**(9), 45–58.

Ben-David, A. (2001). Language acquisition and phonological theory: Universal and variable processes across children and across languages. Unpublished PhD dissertation, Tel-Aviv University, Israel. [in Hebrew].

Berman, R. A. (2001). New perspectives in lexical categorization. Paper presented at the conference on Lexical and Morphological Processing in Spoken and Written Language, Bar Ilan University.

Bolozky, S. (1982). Remarks on rhythmic stress in Modern Hebrew. *Linguistics* **18**, 275–89.

Brent, M. R. & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition* **81**, B33–B44.

Christophe, A., Guasti, T., Nespor, M., Dupoux, E. & Ooyen, B. V. (1997). Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition. *Language and Cognition* **12**, 585–612.

Cohen-Gross, D. (1987). The morphological-syllabic structure of Modern Hebrew. Unpublished PhD dissertation, Bar-Ilan university, Israel [in Hebrew].

Cole, R. & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (ed.), *Perception and production of fluent speech*, 133–63. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language* **2**, 133–42.

Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* **14**, 113–21.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

Echols, C. H., Crowhurst, M. J. & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language* **36**, 202–25.
Echols, C. H. & Newport, E. L. (1992). The role of stress and position in determining first words. *Language Acquisition* **2**, 189–220.
Gleitman, L. R., Newport, E. L. & Gleitman, H. (1984). The current status of motherese hypothesis. *Journal of Child Language* **11**, 43–79.
Gleitman, L. R. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (eds), *Language acquisition: The state of the art*, 3–48. Cambridge: Cambridge University Press.
Gout, A., Christophe, A. & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language* **51**, 548–67.
Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R. & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review* **7**, 504–09.
Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* **44**, 458–567.
Jusczyk, P. W., Cutler, A. & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development* **64**, 675–87.
Jusczyk, P. W., Hohne, E. A. & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics* **61**, 1465–76.
Jusczyk, P. W., Houston, D. M. & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology* **39**, 159–207.
Kelly, M. H. & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception & Performance* **14**, 389–403.
Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
MacWhinney, B. (2008). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
Mattys, S. (2000). The perception of primary and secondary stress in English. *Perception and Psychophysics* **62**, 253–65.
Mattys, S. L. & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition* **78**, 91–121.
Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, N., Bertoncini, J. & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition* **29**, 143–78.
Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S. & Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech* **48**, 279–98.
Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S. & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for cross-linguistic differences. *Journal of Memory and Language* **54**, 283–99.
Nir-Sagiv, B. (2005). Crosslinguistic and developmental perspective on word length as a criterion for vocabulary complexity. Paper presented at the 10th International Congress of the International Association of the Study of Child Language (IASCL), Berlin.
Peretz, I., Lussier, I. & Béland, R. (1998). The differential role of syllabic structure in stem completion for French and English. *European Journal of Cognitive Psychology* **10**, 75–112.
Polka, L. & Sundara, M. (2003). Word segmentation in monolingual and bilingual infant learners of English and French. *In Proceedings of the 15th International Congress of Phonetic Sciences*, 1021–24. Barcelona, Spain.
Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* **274**, 1926–28.
Segal, O. (in progress). The influence of language on the preference and discrimination of stress patterns by infants 9–14 months old. PhD dissertation, Tel-Aviv University, Israel.
Shi, R., Morgan, J. L. & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross linguistic perspective. *Journal of Child Language* **25**, 169–201.
Shi, R., Werker, J. F. & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition* **72**, B11–B21.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press

Spitzer, S. M., Liss, J. M. & Mattys, S. L. (2007). Acoustic cues to lexical segmentation: A study of resynthesized speech. *The Journal of the Acoustical Society of America* **122**, 3678–87.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Woodward, J. Z. & Aslin, R. N. (1990). Segmentation cues in maternal speech to infants. Paper presented at the 7th biennial meeting of the International Conference of Infants Studies, Montreal, Québec, Canada.

https://doi.org/10.1017/S030500090800915X Published online by Cambridge University Press