Word frequencies of preschoolers' oral speech: an analysis on Greek Speaking Children Corpus (GSCC)

Elina Chadjipapa
Democritus University of Thrace
elinaxp@hotmail.com

Περίληψη

Στόχος της παρούσας έρευνας είναι η καταγραφή των συχνοτήτων εμφάνισης λέξεων στο Greek Speaking Children Corpus (GSCC), σώματος κειμένου που περιλαμβάνει προφορικά κείμενα παιδιών προσχολικής ηλικίας. Η έρευνα εστιάζει στη συχνότητα εμφάνισης γραμματικών κατηγοριών με σκοπό τη διερεύνηση της γραμματικής δομής του παιδικού λόγου και την πιθανή επίδραση του φύλου. Οι ποσοτικές αναλύσεις έδειξαν ότι τα παιδιά ηλικίας 3-6 χρονών χρησιμοποιούν με μεγαλύτερη συχνότητα τις γραμματικές κατηγορίες των ρημάτων, των ουσιαστικών και αυτή των άρθρων σε αντίθεση με τα μικρά ποσοστά συχνότητας που παρατηρήθηκαν στις κατηγορίες των επιθέτων και των προθέσεων. Η εξέταση των συχνοτήτων εμφάνισης λέξεων μπορεί να συμβάλλει στην αποτελεσματικότερη διδασκαλία του λεξιλογίου είτε της Γ1 είτε της Γ2.

Λέξεις-κλειδιά: Σώμα κειμένων Προφορικού λόγου Παιδιών Προσχολικής Ηλικίας, συχνότητα εμφάνισης λέξεων, γραμματικές κατηγορίες, προσχολική ηλικία.

1 Introduction

Vocabulary has been at the centre of attention in many researches that have investigated vocabulary acquisition or vocabulary learning. Nation (1993) refers to a sustained interaction between vocabulary knowledge and language use, as the increasing vocabulary size helps learners to improve the language use. Vocabulary knowledge is not the only prerequisite for increasing language skills, but it is a factor that enables the learner to develop receptive and productive language skills.

The perspective that vocabulary size reflects the educational level or someone's productive language skills, evoke the measurement of native speakers' vocabulary size. Researchers suggest that for languages with vocabulary size around 20.000 word families is expected that native speakers add approximately 1.000 words families per year (Nation 1997). That means that a pre-schooler has a vocabulary size of around 3.000 to 4.000-word families when (s)he begins school. Although, a diverse range of theories has been proposed and the increased rate of lexical acquisition is differently estimated (Clark 1993, Nagy 1997). Still vocabulary size is clearly a central problem, as these numbers cannot be adopted in all languages.

Most theories appear to assume the positive role that frequency has in language use. Carol (2008: 103) mentions that among other factors word frequency is the major factor that influences the process of lexical access -*the process by which we activate our word knowledge*- or the process of retrieving lexical information from memory. Nation (1997: 9) considers word frequency as the factor that measures *"how often the word occurs in normal use of the language"*. In speech some words occur more often (high-frequency words) than others (low-frequency words); consequently, the more high-frequency words a learner knows the larger proportion of running words in a written or spoken text the learner comprehends.

2 Research Background

2.1 Word frequency

Research in word frequency appears to have a general consensus on the positive effect of frequency on vocabulary acquisition, language learning and memory performance; the greater the frequency with which a word is produced, the earlier it will be learned. Most of the studies investigate the effect of frequency in vocabulary learning.

Many of them have shown that frequency has an effect on individual words by examining the overall vocabulary size. It is long standing that children acquire new words more quickly when more input is provided by their parents (Huttenlocher et al. 1991; Goodman et al. 2008; Li and Fang 2011). Although, some studies investigated frequency on small sets of novel words (Schwartz and Terrell 1983). The majority of the studies examined the frequency effects on monolingual children in typical development. However, some attention has been paid to how frequency affects children with specific language impairment (Rice et al. 1994) and second-language learners (Wang and Koda 2005).

A considerable number of studies have investigated how frequency affects the children's preferences in word classes. English speaking children seem to use a large proportion of nouns in their early vocabularies (Gentner 1982; Fenson et al. 1994), but on the other hand, for other languages, such as Korean and Chinese, children use verbs more frequently in their early vocabularies. Concerning the grammatical category in terms of word classes a sufficient number of studies have been limited to only a single lexical category such as verbs (Naigles and Hoff-Ginsberg 1998; McDonough et al. 2011; Ashkenazi et al. 2016) and adjectives (Blackwell 2005).

Although, several studies have investigated word frequency and suggest it as an important factor most of them have been restricted to written language. Some of them were directed to child oral speech and few of them were based on oral corpora.

2.2 The effect of gender

An increasing consensus of findings that considers language use as a social phenomenon suggests that both genders use language differently. Conversely, a number of theorists have argued against the existence of any meaningful differences between the two genders (Weatherall 2002). Studies on gender in language learning theory have shown that male and female can reach high levels of language proficiency (Nyikos 2008) but they process information and acquire knowledge differently (Nyikos 1990).

Other studies have examined the actual words people use and the findings have shown that females use more intensive adverbs, more conjunctions, modal auxiliary verbs and first-person singular (Mulac and Lundell 1986; Mulac et al. 2001; Mehl and Pennebaker 2003). On the other hand, male use longer words, more articles, and references to location (Mulac and Lundell 1986; Mehl and Pennebaker 2003). Function words, such as pronouns, are used at much higher rates in conversation, especially by women and men elected to talk about concrete objects, which require nouns and articles.

3 Study

3.1 Aims of the study

Taking advantage of the Greek Speaking Children Corpus (GSCC) the present study investigates the word frequency patterns of word classes in order to explore the composition of child language in terms of word classes. The word classes that were examined are nouns, verbs, adjectives, adverbs, articles, prepositions, pronouns and conjunctions. The further aim of the study is to explore the effect of gender on the grammatical classes. Girls expected to demonstrate significantly higher means in total and by word class.

3.2 Methods of the study

The updated version of the GSCC was the main source of data. The first version of the GSCC (Chadjipapa 2005) included 151,380 tokens and the audio recordings were approximately calculated in 45 hours of oral speech. The new data included 90,204 words and 16 hours of oral speech. Thus, the data set for the current study consists of 238.463 tokens and 61 hours of audio recorded oral speech. The corpus consists of transcription of the audio recordings of 120 Greek-speaking children between the ages of 3-6.

The children were audiotaped at home or at school classes. The interviews were unstructured in order to record spontaneous speech production. The children were asked to tell a story or a fairy tale or to describe a movie or a game or to narrate daily routines at school or at home. The questions intended to elicit a dialogue and let children speak freely about what they wanted. All cases were Adult–Child conversation. Some of them were recorded at children's homes by their mothers or students of the Democritus University of Thrace who were related to the children. All interviews took place after the parents' consent.

The collection includes samples of children speaking standard Greek (N=107) from various Greek regions or the Cypriot dialect (N=13). In terms of gender corpus includes 68 female and 52 male. Finally, the version of the GSCC that we used includes in total 233.604 tokens (the number includes only words): non-words such as laughter or hesitation phenomena and 16.495 word types were not calculated.

3.3 Procedure

Firstly, a sub-corpus including only the child language was extracted. Then, in order to research the major word classes, such as nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions and interjections, a linguistic platform was used. Nooj[1] is a functional environment for a number of linguistic applications and it can process texts or corpora in real and local grammars. The Greek Nooj module which contains the Greek Nooj dictionary and the Greek inflectional grammar (Gavriilidou et al. 2008; Papadopoulou and Chadjipapa forthcoming) were applied as resources in order to perform an extensive linguistic analysis on each individual text in all archives. However, there was still the need to process some of the data manually. That means that the "unknown words" had to be grammatically categorized manually. With the term "unknown word" we refer to the words that were not included in the Greek

[1] For more information see:
http://www.nooj-association.org/index.php?option=com_content&view=featured&Itemid=464

Nooj dictionary and could be categorized in three groups. The first group includes words that were not pronounced correctly (e.g μπολώ/ μπορώ; κούφω/ σκούφο, etc.), the second group refers to words that were invented by the children (e.g. ψευτιάρα/ψεύτρα 'liar'; αερόφωνο/μικρόφωνο 'microphone'), νυχτίδιο/ νυχτερίδα 'bat', etc.), and the last group contains loanwords (e.g. glitter, tablet, etc.). Two groups of the unknown words, children, invented forms and loanwords, were not grammatically annotated and treated together as an individual group, named "Others". In the same group proper nouns were included.

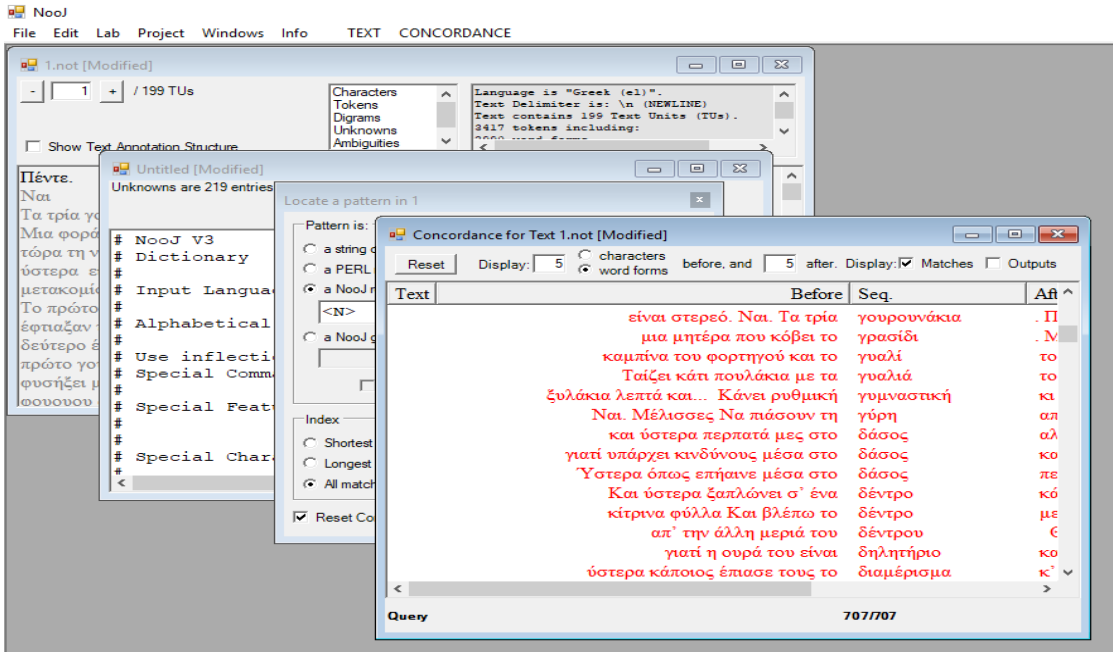

Figure 1 | The analysis of the GSCC in the platform Nooj

The microstructure of the Greek Nooj dictionary provides more coded information concerning the grammatical category of the words besides the conventional part of speech that are investigated in the current study. Since these parts of speech have very few tokens, they were classified with the unknown words in the group "Others".

After this, the Statistical Package for the Social Science (PASW statistics 18) was applied, to check if there is a correlation between the children's word frequency and gender.

4 Results

4.1 Word classes in child language

The type and token information for each word class in the children's language is shown in Table 1. In terms of vocabulary use, children use many more tokens of verbs, nouns; articles[2] and conjunctions than any other word class (see Table 1). They seem to use more verbs than nouns.

[2] Articles and clitic forms of the pronouns are not disambiguated.

Word Class	Type	Type proportion (type/ total type%)	Token	Token proportion (token/ total token%)
noun	1625	42,48%	16.255	13,25%
verb	661	17,29%	18.764	15,29%
adjective	373	9,76%	4.980	4,06%
adverb	153	4,00%	10.744	8,76%
article	20	0,52%	14.728	12,00%
preposition	15	0,39%	2.752	2,24%
pronoun	44	1,15%	7.133	5,81%
conjunction	36	0,94%	14,383	11,72%
interjection	21	0,55%	714	0,58%
other	875	22,89%	32.261	26,29%
Total	3.823	100%	122.714	100%

Table 1 | Type and token information for each word class of child language

Moreover, children use many more adverbs and pronouns than adjectives, prepositions and interjections as shown in Figure 2.

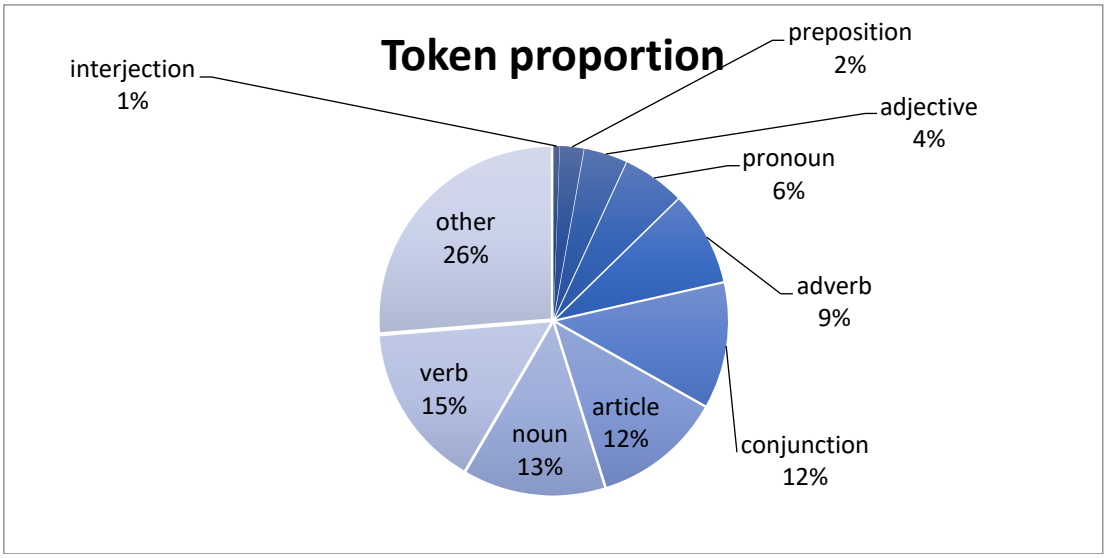

Figure 2 | Token proportion of major word classes in child language

However, regarding the word type total as children's vocabulary size, the linguistic analysis provides a different composition in terms of word classes in children's language use (see Figure 2). The results showed that nouns, verbs and adjectives occupy the major parts of their vocabulary, contrary to the other word classes. Still, nouns and verbs seem to be acquired more than the other classes with the difference that nouns are used more than verbs (see sec. 5 Discussion).

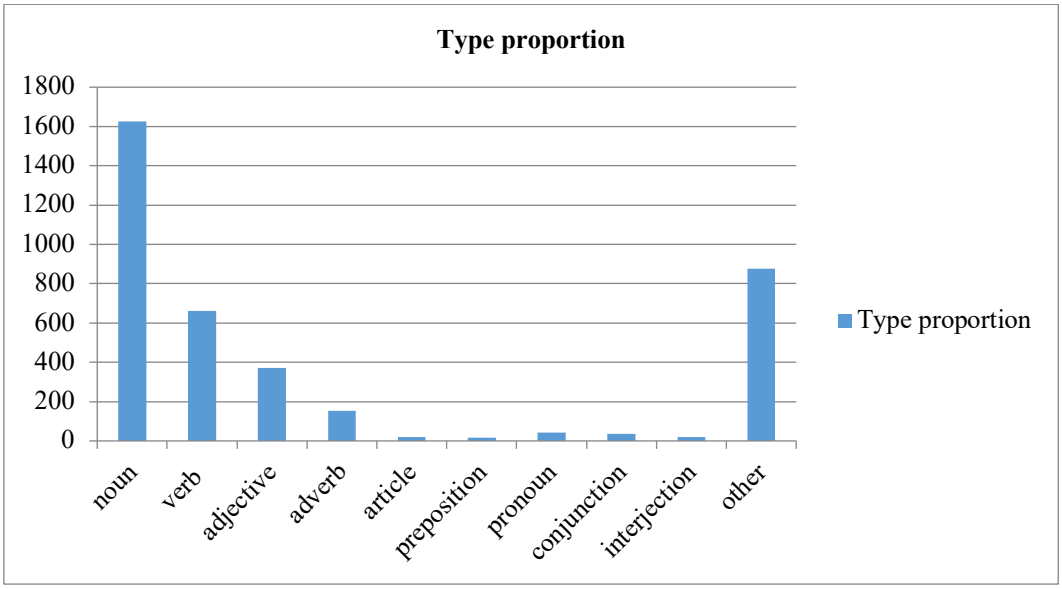

Figure 3 | Type proportion of children's vocabulary

Further analysis highlights that the most frequent words for each word class was found to be a) the conjunction *και;* b) the article *το* and *τα*; c) the verb *είμαι*; d) the noun *μαμά,* and e) the adverb *ναι*.

4.2 The effect of Gender on word frequency

Descriptive statistics were used for investigating the effects of gender in word frequency in child language in terms of word classes. T-test was used to investigate if gender affects a) the number of tokens that children use and b) the frequency use of the major word class.

The main effect of gender on tokens number and on in each separate word class was found (statistically) insignificant. Boys and girls appear to use the same number of tokens in total and for most of the word major classes. Table 2 reports the mean scores of each word class by gender. The analysis indicates that gender affects the particular word class.

Gender		Article	Pron.	Conj.	Noun	Verb	Adj.	Prep.	Adv.	Interj.
Female	**M**	140,79	68,51	135,56	148,38	165,29	44,82	23,76	96,38	6,50
	SD	97,961	94,062	93,331	96,248	98,252	28,514	16,630	52,099	7,939
Male	**M**	99,12	47,58	99,33	118,56	144,69	37,15	21,85	80,58	5,23
	SD	74,961	34,105	80,427	85,624	101,133	30,970	22,574	54,805	6,629

Table 2 | Mean scores of word classes by gender

5 Discussion

The purpose of the present study was to investigate the word frequency patterns of the major word classes such as nouns, verbs, adjectives, adverbs, articles, prepositions, pronouns, conjunctions and interjections in child language. The GSCC was analyzed and the linguistic platform Nooj including the Greek Nooj version was used, in order to explore the composition of child language in terms of word classes.

It was found that the most frequently used word classes calculated in tokens are verbs, nouns, articles and conjunctions. However, the most frequent word class as concerns the word types is the one of nouns. The results designate a difference between nouns and verbs. In other words, verb tokens outnumber noun tokens, but concerning the word types that children use, nouns outnumber verbs. That could be due to the heavy inflectional system of the Greek language and the variety of the verbs' inflectional forms contrary to nouns' (e.g., for verbs: person, number, tense, voice etc.; for nouns: case, number) and explains to some extend the difference between tokens and types of the two word classes. Gentner (1982) argues that verbs present a relatively small number of types with a high token frequency. Concerning the Greek inflectional system, word types is the unit that reveals the diversity in Greek children's vocabulary use in terms of word classes and denotes that nouns are used more than verbs.

The findings of the study are in line with other researches (Nelson 1973; Gentner 1982; Li and Fang 2011) which provide evidence that nouns tend to be acquired earlier than verbs. Some researchers asserted that nouns are more concrete and imageable than verbs which tend to attract relatively low imageability ratings (Gentner and Boroditsky 2001; Gentner 2006; Ma et al. 2009; McDonough et al. 2011). The "noun bias" seems to be a universal phenomenon since objects are more stable than concepts of actions or events (Gentner 1982) and because children hear more nouns than other word classes (Sandhofer et al. 2000: 562). For the same reason children tend to use far fewer pronouns, adjectives, prepositions and interjections. These word classes are more abstract and less imageable than nouns, which have concrete referents.

The final aim was to analyze the effect of gender on the choice of grammatical classes and the differences between their frequencies. It was expected that girls would demonstrate higher means in total token and types and in all word classes. This hypothesis was not supported by the data. The results of the study indicated that male and female use equally all word classes. To our knowledge, there are few studies that investigate gender as a factor that influences word frequency patterns of the word classes in child language. However, previous research evidence has indicated that females perform better in vocabulary learning (Newman et al. 2008) but there is small accounting of the variance (Fenson et al. 1994). This result could be attributed to the fact that the effect of gender interrelates with age or educational level.

6 Conclusions and Further research

The present study investigates the word frequency in children's vocabulary in terms of word class by analyzing a corpus of oral spontaneous children speech. It was expected to find that children tend to use many more words with concrete referents such as nouns than less concrete words such as adjectives, adverbs, prepositions, interjections and conjunctions. This hypothesis was confirmed by the data. In addition, children seem to acquire monosyllabic words more easily and use them much more frequently than multi-syllabic words such as "*μαμά*".

As concerns, the effects of gender on word frequencies, an insignificant interaction was found. Gender does not seem to influence the use of all word classes. The results raise the necessity of the extension of the GSCC, in order to check the gender's influence in vocabulary use and acquisition.

However, the data requires further analysis regarding the GSCC. Proper nouns and loanwords will provide more information about the word frequencies and the children's language. Also, a comparison between children's and adult's speech will reveal the relation between linguistic input and children's first word uses. Finally, the correlation between word frequencies, neighbourhood density, age, Type Token Ratio and Mean Length Utterance will give more information about the children's vocabulary use. Profiling Greek pre-schoolers in vocabulary use can contribute to ameliorate the curricula for Greek Language by creating word lists per age for typical development children or for children with language impairments and to second language teaching.

References

Ashkenazi, Orit, Ravid, Dorit, and Steven Gillis. 2016. "Breaking into the Hebrew Verb System: A Learning Problem." *First Language* 36:505–524.

Blackwell, Aleka A. 2005. "Acquiring the English Adjective Lexicon: Relationships with Input Properties and Adjectival Semantic Typology." *Journal of Child Language* 32:535–562.

Carroll, Carroll W. 2008. *Psychology of language*. Fifth edition. Belmont, CA: Cengage Learning/Wadsworth.

Chadjipapa, Elina. 2005. "Construction of the Greek Speaking Children Corpus." [IN GREEK] Unpublished senior thesis, Democritus University of Thrace.

Clark, Eve V. 1993. "The Lexicon in Acquisition." *Cambridge Studies in Linguistics Vol. 65.* Cambridge: Cambridge University Press.

Fenson, Larry, Dale, S. Philip, Reznick, Steven, Bates, Elizabeth, Thal, Donna, Stephen J. Pethick, Tomasello, Michael. Carolyn, B. Mervis, and Joan Stiles. 1994. "Variability in Early Communicative Development." *Monographs of the Society for Research in Child Development 59 (5)* Serial No. 242.

Gavriilidou, Zoe, Papadopoulou, Eleni, and Elina Chadjipapa. 2008. "The New Greek NooJ Module: Morphosemantic Issues." In *Proceedings of the 2007 NooJ International Conference*, edited by Blanco, X., and M. Silberztein, 96-103. Cambridge: Cambridge Scholars Publishing.

Gentner, Dedre. 2006. "Why Verbs are Hard to Learn." In *Action Meets Word: How Children Learn Verbs*, edited by Hirsh-Pasek, K., and R. Golinkoff, 544-564. Oxford University Press.

Gentner, Dedre, and Lera Boroditsky. 2001. "Individuation, Relativity, and Early Word Learning." In *Language Acquisition and Conceptual Development*, edited by Bowerman, M., and S. Levinson, 215-256. New York: Cambridge University Press.

Goodman, Judith, Dale, Philip, and Ping Li. 2008. "Does Frequency Count? Parental Input and the Acquisition of Vocabulary." *Journal of Child Language* 35: 515–531.

Huttenlocher, Janellen, Haight, Wendy, Bryk, Anthony, Seltzer, Michael, and Thomas Lyons. 1991. "Early Vocabulary Growth: Relation to Language Input and Gender." *Developmental Psychology* 27:236–248.

Li, Hanhong, and Alexa C. Fang. 2011. "Word Frequency of CHILDES Corpus: Another Perspective of Child Language Features." *International Computer Archive of Modern and Medieval English* 35:95–116.

Nagy, William E. 1997. “On the Role of Context in First- and Second-language Vocabulary Learning.” In *Vocabulary: Description, Acquisition, and Pedagogy*, edited by N. Schmitt and M. McCarthy, 64-83. Cambridge: Cambridge University Press.

Naigles, Letitia. R., and Erika Hoff-Ginsberg. 1998. “Why are some Verbs Learned Before Other Verbs? Effects of Input Frequency and Structure on Children's Early Verb Use.” *Journal of Child Language* 25:95–120.

Nation, I. S. Paul. 1993. “Vocabulary Size, Growth and Use.” In *The Bilingual Lexicon*, edited by R. Schreuder and B. Weltens, 115-134. Amsterdam/Philadelphia: John Benjamins.

Nation, I. S. Paul. and Robert Waring. 1997. “Vocabulary Size, Text Coverage and Word Lists.” In *Vocabulary: Description, Acquisition and Pedagogy*, edited by Schmitt, N. and M. McCarthy, 6-19. Cambridge: Cambridge University Press.

Nelson, Katherine. 1973. “Structure and Strategy in Learning to Talk.” *Monographs of the Society for Research in Child Development* Vol. 38, Nos 1–2, Serial No. 149:1–135.

Nyikos, Martha. 1990. “Sex-related Differences in Adult Language Learning: Socialisation and Memory Factors.” *Modern Language Journal* 74 (3):273-287.

Nyikos, Martha. 2008. “Gender in Language Learning.” In *Lessons from Good Language Learners: Insights for Teachers and Learners*, edited by C. Griffiths, 73-82. Cambridge University Press.

Ma, Weiyi, Golinkoff, Roberta M., Hirsh-Pasek, Kathy, McDonough, Colleen, and Twila Tardif. 2009. “Imageability Predicts the Age of Acquisition of Verbs in Chinese Children.” *Journal of Child Language* 36:405–423.

McDonough, Colleen., Song, Lulu, Hirsh-Pasek, Kathy, Golinkoff, Roberta M., and Robert Lannon. 2011. “An Image is Worth a Thousand Words: Why Nouns Tend to Dominate Verbs in Early Word Learning.” *Developmental Science*, 181–189.

Mehl, Matthias R., and James W. Pennebaker. 2003. “The Sounds of Social Life: A Psychometric Analysis of Students’ Daily Social Environments and Natural Conversations.” *Journal of Personality and Social Psychology* 84:857–870.

Mulac, Anthony, Bradac, James J., and Pamela Gibbons. 2001. “Empirical Support for the Gender-as-culture Hypothesis: An Intercultural Analysis of Male/Female Language Differences.” *Human Communication Research* 27:121–152.

Mulac, Anthony, and Torborg Louisa Lundell. 1986. “Linguistic Contributors to the Gender-linked Language Effect.” *Journal of Language and Social Psychology* 5:81–101.

Newman, Matthew L., Groom, Carla J., Handelman, Lori D., and James W. Pennebaker. 2008. “Gender Differences in Language Use: An Analysis of 14.000 Text Samples.” *Discourse Processes* 45 (3):211–236.

Papadopoulou, Eleni, and Elina Chadjipapa. Forthcoming. “A Morphological Grammar for Modern Greek: State of Art, Evaluation and Upgrade.” *Proceedings of the 2020 NooJ International Conference*, Cambridge: Cambridge Scholars Publishing.

Rice, Mabel L., Oetting, Janna B. Marquis, Bode, Janet, and Soyeong Pae. 1994. “Frequency of Input Effects on Word Comprehension of Children with Specific Language Impairment.” *Journal of Speech and Hearing Research* 37:106–121.

Sandhofer, Catherine M., Linda B. Smith, and Jun Luo. 2000. “Counting Nouns and Verbs in the Input: Differential Frequencies, Different Kinds of Learning?” *Journal of Child Language* 27:561–585.

Schwartz, Richard. G., and Brenda. Y. Terrell. 1983. “The Role of Input Frequency in Lexical Acquisition.” *Journal of Child Language* 10:57–64.
Wang, Min., and Keiko Koda. 2005. “Commonalities and Differences in Word Identification Skills Among Learners of English as a Second Language.” *Language Learning* 55:71–98.
Weatherall, Ann. 2002. *Gender, Language, and Discourse*. London: Routledge.