Kernel Vocabulary and Zipf's Law in Maternal Input to Syntactic Development

Anat Ninio
The Hebrew University

1. Power-law distribution of form-class items

It is a very old finding, and it has the status of a universal law, that if we take a very large text and rank the words in order of their frequency of use, we get a severely skewed distribution, with a few very frequent items and very many infrequent ones. Formally, when the frequency of words is plotted as a function of their rank-order in the vocabulary, the resultant graph follows a power law. This is the Zipf-Mandelbrot law (Mandelbrot, 1966; Zipf, 1935/1965) that actually applies to all sorts of large sets of items that form a complex system, but it was very early on established regarding the behavior of words used in a text.

In the reported study the hypothesis was tested that the same would be true of verbs appearing in specific types of syntactic word combinations, namely, belonging to a specific form-class. A form-class is a subgroup of the total lexicon with similar grammatical behaviour. Such a group of items is a priori quite homogeneous as the items are selected for possessing identical syntax, so it is not trivial to expect that they, too, would possess a very skewed power-law distribution.

We were in particular interested in the global statistical features of the so-called "motherese" speech register, namely, the speech of mothers addressed to young children, which is the main input to the acquisition process. Previous studies showed that maternal input has a skewed distribution, so that some verbs account for a very high percentage of all utterance tokens in various syntactic patterns. For example, Goldberg, Casenhiser and Sethuraman (2004) analyzed English-speaking mothers' utterance tokens in a large corpus and concluded that in different argument structure constructions, among them the Subject-Verb-Oblique Intransitive and the Subject-Verb-Object-Object2 Ditransitive frames, one specific verb accounts for a large percentage of utterance tokens. Similar skewed distributions in English maternal speech samples were reported by Naigles and Hoff-Ginsberg (1995), Sethuraman and Goodman (2004) and Theakston, Lieven, Pine and Rowland (2004), and for Hebrew maternal samples by Ninio (1999a) and Ninio (1999b).

Previous studies, however, did not go beyond estimating the relative frequency of a few items from the total distribution. In the reported study, we plotted the whole range of items used by Hebrew-speaking mothers in one

specific syntactic construction, and tested the hypothesis that the distribution will obey a power law.

The focus of the study are maternal input sentences containing a verb or adjective followed by an indirect object with a *le-* preposition. Like the English `*to'*, the Hebrew *le-* has a homophone used as an adverb of direction; sentences with the adverbial use were not included in the study.

The maternal speech sample used in the study was the pooled corpus of 48 Hebrew-speaking mothers talking with their young children who were between 0;10 and 2;8 at the time of the observations. The speech samples were taken from a videotaped observational study (Ninio, 1984). There are more than 56,000 multiword utterances in the pooled multiword corpus. The corpus was searched for sentences in which there was a verb or adjective followed by an indirect object with a *le-* preposition. Overall, there were 6956 utterances of this kind, representing 230 different verbs and adjectives.

Figure 1. presents the rank/frequency distribution of all the verbs appearing with an indirect object in the pooled corpus of 48 mothers.
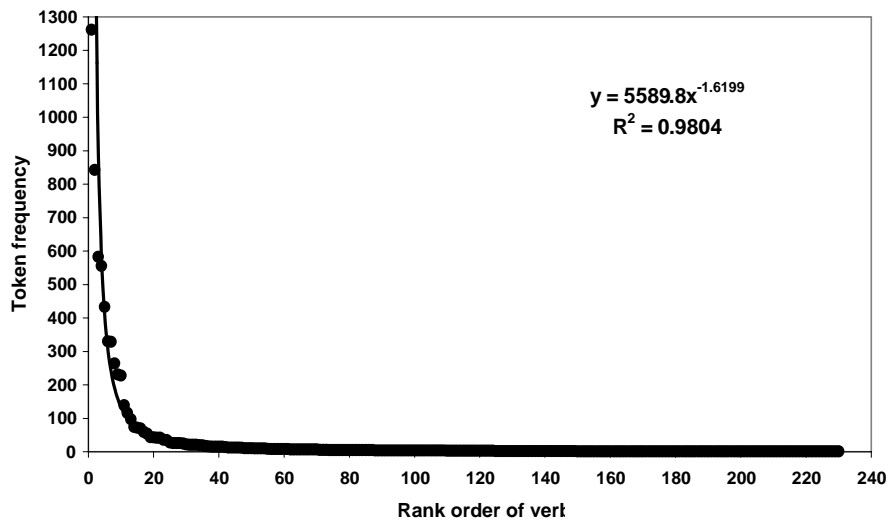

The chart shows a power-law distribution with the fitted equation:

$$y = 5589.8x^{-1.6199}$$
$$R^2 = 0.9804$$

with axes labeled "Token frequency" (vertical) and "Rank order of verb" (horizontal).

Figure 1. Rank-frequency distribution of maternal VI sentences with fitted power-law Zipf curve

We can observe the very clear power-law distribution of the 230 verbs; the fit of the power-law curve is excellent (98%). We might conclude that the use frequency of verbs participating in the indirect-object construction has a typical Zipf distribution.

There is another way to present the same statistical feature and that's the Pareto presentation. Mathematically the two are transformations of each other, but graphically they turn the axes around.
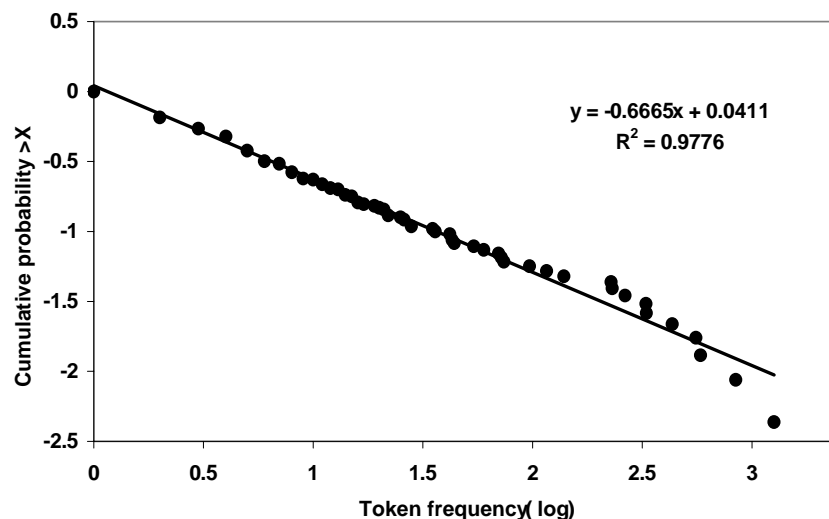

$$y = -0.6665x + 0.0411$$
$$R^2 = 0.9776$$

Figure 2. Cumulative probability of maternal VI sentences having larger token frequency than X, with fitted Pareto function (log-log plot)

On the X axis we plot the token frequency of items ordered by rank, and it increases as we get higher values; on the Y axis we plot the cumulative probability of an item having equal or larger frequency than X. For the lowest value, the probability is 100%, and it decreases as the frequencies get higher. The cumulative probabilities also distribute under a power-law function, but Figure 2 presents them on a log-log plot, in which the power-law distribution shows up as a straight line.

The fit is like it was in the Zipf presentation we saw before, namely, 98%. However, in the Pareto graph we can see very clearly that the fit of the curve is not so good when we come to the highest values.

2. Kernel vocabulary

Zipf pointed out that Zipf's Law does not apply to the very-frequent items at the start of the distribution. The exceptions to Zipf's Law can be exploited for a closer examination of the maternal input. Ferrer Cancho and Sole (2001) and Montemurro (2001) systematically explored the exceptions in general corpora

of English texts and found that the total vocabulary can be divided into two or more power-law registers, differing in their mathematical distribution as well as their content. They showed that regardless of sample size, there is a set of very-frequent items that have a less steep decay exponent in the power-law distribution of rank by frequency than other items. The crucial finding was that **the quantitative difference is also a qualitative one.** The registers represent two kinds of vocabulary items: the very-frequent items belong to a basic vocabulary, while the less-frequent items are specific words. Their estimate of the basic vocabulary of English is about 4,000-6,000 words.

We wanted to use this method to identify the nuclear items of verb sub-categories.

The idea that classes of verbs have nuclear items comes from Dixon (1982, pp. 121-125) who claims that universally, all open classes in the lexicon of natural languages contain subsets each of which consist of one or very few generic items and a larger set of more specific items which are their almost-hyponyms. The generics share the semantics and syntax of their more specific relatives; however, they have more extensive and more varied semantic and syntactic options than the more specific items.

This analysis was repeated on the indirect-object construction in Hebrew maternal speech. Figure 3 presents the Pareto graph, but this time the set has been cut into 5 domains of 10 values each.
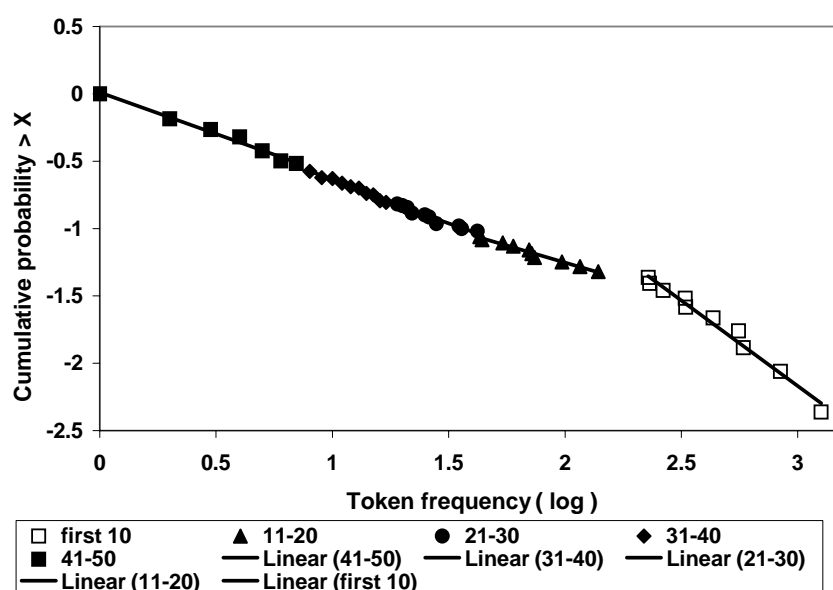

Figure 3. Pareto graph: Cumulative probability of maternal VI sentences having larger token frequency than X, for 6 domains (log-log plot)

All but the first domain of 10 items had the same distribution; the first 10 were different. We may conclude that there are two different domains, shown in Figure 4:

$$y = -0.6432x + 0.0088$$
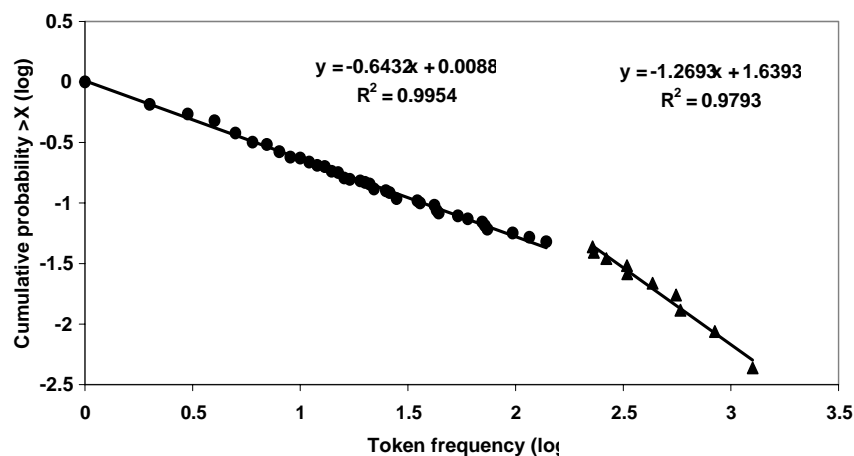$$R^2 = 0.9954$$

$$y = -1.2693x + 1.6393$$
$$R^2 = 0.9793$$

Figure 4. Cumulative probability of maternal VI sentences having larger token frequency than X, with fitted Pareto function for 2 regimes (log-log plot)

The first 10 most frequent items -- all verbs -- form a separate register with a different power-law exponent, whereas all less-frequent items have the same exponent. We can see this phenomenon best if we return to the Zipf curve of Figure 1, this time separating out the two power-law regimes. Figure 5 presents the graphs.
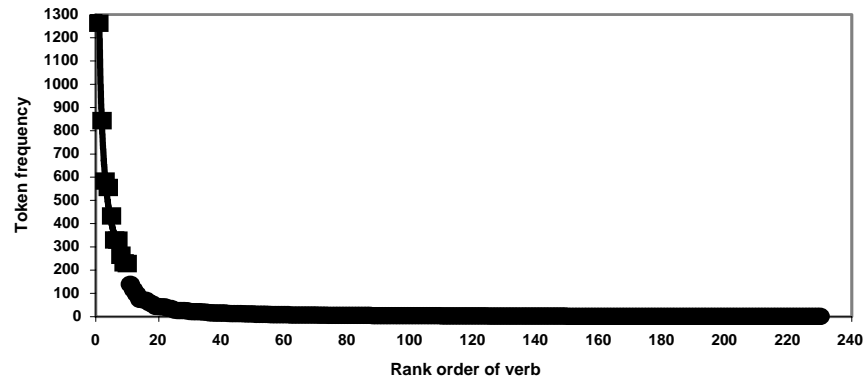
Figure 5. Rank-frequency distribution of maternal VI sentences with fitted power-law Zipf curves for 2 regimes

Namely, the first part of the Zipf curve in Figure 1 is in fact under a separate power-law regime. The curves of both regimes have a very high fit (98%).

The quantitative analysis thus identifies 10 verbs as the kernel vocabulary for this construction in the maternal input. The next question we turn to is, what are these verbs and why are there 10 of them?

3. The kernel vocabulary for the Hebrew ditransitive verbs

We used the method of identifying two or more different power-law regimes in order to point to the verbs occupying the different regimes. The strategy separated the vocabulary into two sets of items that behave quantitatively differently. Now we want to see what are the characteristics of the two groups and whether we repeat the finding reported by Zipf and by Ferrar and Sole, namely, that the most frequent items in a Zipf-distribution are semantically very general. Table 1 presents the 10 most frequent verbs in the VI pattern.

Table 1. First 10 most frequent verbs in the VI pattern:

Hebrew verb	Translation equivalent
NATAN	'give'
HAYA	'has'
HEVI	'bring'
ASA	'make, do'
HERA	'show'
SIPER	'tell'
HIGID	'say'
AZAR	'help'
SAM	'put'
QARA	'name'

Two phenomena are evident: first, all the verbs are very general.

Second, they cover a wide variety of semantics. The verbs belong to at least 6 different semantic fields: Transfer of objects (*give, bring, put);* Possession *(have,)*; Creation of effected object *(make/do)*; Transfer of information (*show, tell story, say*), Provision of service (*help),* and Refer (*name*).

As an example, Table 2 presents all the verbs that appeared in the maternal sentences with a VI pattern whose meaning with an indirect object fell into the semantic field of "Provide a Service", as in, for example, "*clean it for me*" or "*open it for me".*

Table 2. Verbs in the semantic field "PROVIDE A SERVICE" when receiving an indirect object

help

change (diaper)	clean
close	cool
count	cut
cut/trim nails	dress (tr)
empty	examine
glue	guard
hold	inflate
light	open
fix	paint/color
peek	put-on-shoe
put-on-sock-tr	read
read-to	screw
take-apart	tie
turn-on	turn-tr
wash	wash-hair
weigh	wipe
wipe-nose	

The generic verb *help* was among the 10 kernel items; It appeared in sentences like "*Let me help you*". The other 32 verbs refer to more specific services that can be provided, such as "*Do you want me to tie it for you*?". Both, in Hebrew, have an identical Indirect-Object manifestation.

Similar analyses can be made for the other kernel verbs; each is a generic verb for one of the semantic subsets of all verbs that occur in the VI pattern.

The most important conclusion for language development is that the kernel vocabulary for the verb/adjective-indirect-object construction consists of a group of verbs with varied semantics. There is no single syntax-semantics mapping in this grammatical pattern in Hebrew; in fact there are at least 8 different semantic categories of verbs taking an indirect object in Hebrew. However, each slice of the syntactic pattern is covered by one or more verb-frequently used, generic verb of its own. Namely, we should be open to the possibility that syntax-semantics linking is a multiple phenomenon even within a single syntactic pattern, and that each of these slices of grammar makes some use of the kernel items which are presented to children with extremely high frequency in the input. Some implications of this and similar phenomena are explored in Ninio (in press).

Finally, these results have implications for understanding syntactic networks (Ferrer, Sole, & Kohler, 2003) as well as the acquisition process. It appears that the network generated by a single form-class (containing the different verbs that can occur with a given type of syntactic complement) repeats on a smaller scale and in a fractal manner the generic-specific structure of the total vocabulary. Generic-specific relations seem to be the key to the structure of the syntactic lexicon at all levels of analysis. This is important information about the structure of language that deserves careful further study.

References

Dixon, Robert M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Ferrer i Cancho, Ramon, & Sole, Ricard V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics, 8*, 165–173.

Ferrer i Cancho, Ramon, Sole, Ricard V. & Kohler, Reinhard (2003). Universality in syntactic dependency networks. SFI Working Paper #03-06-042.

Goldberg, Adele E., Casenhiser, Devin & Sethuraman, Nitya (2004). Learning argument structure generalizations. *Cognitive Linguistics, 15,* 289-316.

Mandelbrot, Benoit (1966). Information theory and psycholinguistics: A theory of words frequencies. In Paul Lazarsfeld and Neil Henry (Eds.), *Readings in mathematical social science* (pp. 350-368). Cambridge, MA: MIT Press.

Montemurro, Marcelo A. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A, 300*, 567–578.

Naigles, Letitia & Hoff-Ginsberg, Erika (1995). Input to verb learning: evidence for the plausibility of syntactic bootstrapping. *Developmental Psychology, 31*, 827-837.

Ninio, Anat (1984). Functions of speech in mother-infant interaction. Final Science Report to the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel.

Ninio, Anat (1999a). Model learning in syntactic development: intransitive verbs. *International Journal of Bilingualism, 3*, 111-31.

Ninio, Anat (1999b). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language, 26*, 619-53.

Ninio, Anat (in press). Valency and the Learning Curve: The acquisition of syntax. Oxford: Oxford University Press.

Sethuraman, Nitya & Goodman, Judith C. (2004). Children's mastery of the transitive construction. Paper presented at the Stanford Child Language Research Forum, April 2004.

Theakston, Anna L., Lieven, Elena V. M., Pine, Julian M., & Rowland, Caroline F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language, 31*, 61-99.

Zipf, George K. (1935/1965). *Psycho-biology of languages*. Cambridge, MA: MIT Press. (first published by Houghton Mifflin).