Psych 215L: Language Acquisition

Lecture 12
Morphosyntax

Computational Problem

Determine that there are grammatical categories like Noun and Verb that behave similarly with respect to their morphology and combinatorial syntax.

Noun = \{penguin, goblin, glitter, cheese\}
Morphology: Nouns can take determiners like “the”
\{(the penguin, the goblin, the glitter, the cheese)\}

Verb = \{swim, dance, flutter, smell\}
Morphology: Verbs can take –ed to indicate past tense
Combinatorial syntax: Verbs can take adverbs that modify them, like “really”
\{really swim, really dance, really flutter, really smell\}

Yang 2010

How do we know when children achieve adult-like knowledge?
“Language use is the composite of linguistic, cognitive and perceptual factors many of which, in the child’s case, are still in development and maturation. It is therefore difficult to draw inferences about the learner’s linguistic knowledge from his linguistic behavior.”

“The pioneering work on child language that soon followed, include those who did not follow the generative approach, also recognized the gap between what the child knows and what the child says...child language be interpreted in terms of adult-like grammatical devices, which has continued to feature prominently in language acquisition.”

Example adult-like grammatical device: Verb categories like Noun and Verb

Yang 2010

How do we know when children achieve adult-like knowledge?
“This tradition has been challenged by the item or usage-based approach to language most clearly represented by Tomasello (1992, 2000a, 2000b, 2003), which reflects a current trend (Bybee 2001, Pierre-humbert 2001, Goldberg 2003, Culicover & Jackendoff 2005, Hay & Baayen 2005, etc.) that emphasizes the storage of specific linguistic forms and constructions at the expense of general combinatorial linguistic principles and overarching points of language variation (Chomsky 1965, 1981).”

Properties used in support of item-based approach:
(1) Use of verb in limited “constructions”
(2) Limited morphology on any given verb
(3) Unbalanced determiner usage (ex: use only “the” with some and only “a/an” with others)
The lack of a formal statistical test for productivity

“So far as we can tell, however, these evidence in support for item-based learning has been presented, and accepted, on the basis of intuitive inspections rather than formal empirical tests. For instance, among the numerous examples from child language, no statistical test was given in the major treatment (Tomasello 1992) where the Verb Island Hypothesis and related ideas about item-based learning are put forward. Specifically, no test has been given to show that the observations above are statistically inconsistent with the expectation of a fully productive grammar, the position that item-based learning opposes. Nor, for that matter, are these observations shown to be consistent with item-based learning…”

Zipf’s law

“Under the so-called Zipf’s law (Zipf 1949), the empirical distributions of words follow a curious pattern: relatively few words are used frequently—very frequently—while most words occur rarely, with many occurring only once in even large samples of texts. More precisely, the frequency of a word tends to be approximately inversely proportional to its rank in frequency.”

\[ f = \frac{c}{r} \]

where \( c \) is some constant

Checking Zipf’s law on the Brown corpus

“It is often the case that we are not concerned with the actual frequencies of words but their probability of occurrence; Zipf’s law makes this estimation simple and accurate.”

\[ p_r = \left( \frac{1}{r} \right)^{\text{H} \text{armonic \ Number} \frac{1}{\text{H} \text{armonic \ Number}} \sum_{r=1}^{N} \frac{N-1}{r} + 1 \]
Checking Zipf’s law on the syntactic rules in the Penn-Treebank corpus

“Since the corpus has been manually annotated with syntactic structures, it is straightforward to extract rules and tally their frequencies. The most frequent rule is “PP→NP NP”, followed by “S→NP VP” again, the Zipf-like pattern.”

The moral of Zipf’s law for productivity analyses

“Claims of item-based learning build on the premise that linguistic productivity entails diversity of usage: the “unevenness” in usage distribution is taken to be evidence against a systematic grammar. The underlying intuition, therefore, appears to be that linguistic combinations might follow something close to a uniform distribution.”

A closer look at determiner usage with nouns (among other types of usage)

“Consider a fully productive rule that combines a determiner and a singular noun, or “DP→D N”, where “D→a|the” and “N→cat|book|desk|...”. We use this rule for its simplicity and for the readily available data for empirical tests but one can easily substitute the rule for “VP→V DP”, “VP→V in Construction”, “Vinflection→Vstem + Person + Number + Tense”. All such cases can be analyzed with the methods provided here.”

Expected determiner usage

“Suppose a linguistic sample contains S determiner-noun pairs, which consist of D and N unique determiners and nouns. (In the present case D = 2 for “a” and “the.”) The full productivity of the DP rule, by definition, means that the two categories combine independently.”

Observation 1

“…nouns (and open class words in general) will follow Zipf’s law…relatively few nouns occur often but many will occur only once—which of course cannot overlap with more than one determiners.”

Observation 2

“…while the combination of D and N is syntactically interchangeable, N’s tend to favor one of the two determiners, a consequence of pragmatics and indeed non-linguistic factors.”

Quantifying productivity

S = # of samples in linguistic data set
D = # of unique determiners
N = # of unique nouns

Overlap: A noun occurs with more than one determiner.

Calculating observed overlap

For each noun n in the data set, determine if it occurs with more than one determiner.

If so, overlap(n) = 1.
If not, overlap(n) = 0.

Observed overlap = \( \frac{\sum_{n=1}^{N} \text{overlap}(n)}{N} \)
Yang 2010

Quantifying productivity

\[ S = \# \text{ of samples in linguistic data set} \]

\[ D = \# \text{ of unique determiners} \]

\[ N = \# \text{ of unique nouns} \]

Calculating expected overlap \([O(D,N,S)]\)

“This requires the calculation of the expected overlap value for each of the \(N\) nouns over all possible compositions of the sample.”

Sum individual expected overlap for each noun (from rank 1 to \(N\)) in the data set, and then divide by \(N\) to get the average expected overlap for all nouns.

Individual noun overlap:

Probability that it is not used with only one determiner.

All the instances where it’s not the case that...

...the noun just didn’t get produced in this data set of \(S\) samples for whatever reason...
Quantifying productivity

S = # of samples in linguistic data set
D = # of unique determiners
N = # of unique nouns

Calculating expected overlap \([O(D,N,S)]\)

Individual noun overlap: Probability that it is not used with only one determiner.

...and the noun was sampled, but favored one determiner exclusively for whatever reason.

Calculating the probability that this noun just didn’t get produced for \(S\) samples

\[(1-pr) = \text{probability that noun with rank } r \text{ didn’t appear for this one trial}
\]
...done \(S\) times (quantity²)

(1) probability of noun \(n_r\) (which appears with frequency \(p_r\)) combining with the \(i\)th determiner (which has its own frequency of appearing in the corpus, \(d_i\)) = \(p_r * d_i = dipr\)

(2) probability of all possible compositions of sample \(S\) where \(n_r\) combines with \(d_i\) only

However frequently noun with rank 1 appeared with whatever determiners =...
Quantifying productivity

S = # of samples in linguistic data set
D = # of unique determiners
N = # of unique nouns

Calculating expected overlap \([O(D,N,S)]\)

Calculating the probability that this noun favored one determiner exclusively

(2) probability of all possible compositions of sample S where \(n_r\) combines with \(d_i\) only

\[ P_r = \frac{\text{number of times } n_r \text{ combined with } d_i}{S} \]

... however frequently noun with rank 2 appeared with whatever determiner + ...

... however frequently noun with rank \(r-1\) appeared with only this one determiner \(d_i\) + ...

... and so on...
Calculating the probability that this noun favored one determiner exclusively

(2) probability of all possible compositions of sample S where \( n_i \) combines with \( d_i \) only

\[
\sum_{k=0}^{N} \binom{D}{k} \binom{N-k}{S-k} \frac{1}{D^S}
\]

... for each of the S samples in the data set

Calculating expected overlap \([O(D,N,S)]\)

Another way to derive \((d_{p_i} + 1 - p_i)^S\)

For each sample (quantity\( S \)), we want the probability of \( \neg \) (picking that noun out when it doesn't come with determiner \( d_i \)). This is \((1 - p_i)(1 - d_{p_i}) = 1 - p_i \cdot d_{p_i} + d_{p_i} \cdot 1 - p_i\).
Calculating the probability that this noun favored one determiner exclusively.

Note: This quantity is also equivalent to the following equation, which calculates $p_{nr}$, is sampled but with $d_{i}$ exclusively directly:

\[
\sum_{j=0}^{S} \binom{S}{j} \left( \frac{p_{nr}}{d_{i}} \right)^{j} \left( 1 - \frac{p_{nr}}{d_{i}} \right)^{S-j}
\]

For each combination of $S$ samples:

- A sample where there are $j$ uses of $nr$ and $S-j$ uses of some other noun

...and this is what we use in the original formula

\[
\sum_{i=0}^{D} \sum_{j=0}^{N} \binom{S}{j} \left( \frac{p_{nr}}{d_{i}} \right)^{j} \left( 1 - \frac{p_{nr}}{d_{i}} \right)^{S-j}
\]
Yang 2010

Determiner usage case study: D = {a, the} only

Data = Adam [Brown], Eve [Brown], Sarah [Brown], Naomi [Sachs], Peter [Bloom?], Nina [Suppes]

Age range across all children: 1;1 – 5;1

Comparison sets

Each individual child

- First 100, 300, and 500 productions from all children to capture earliest stage of language production which should (presumably) be the least productive vs.

Adult production estimates from the Brown corpus

---

Yang 2010

Determiner usage case study: D = {a, the} only

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sample Size (S)</th>
<th>a or the/Nouns</th>
<th>Overlap (expected)</th>
<th>Overlap (expected)</th>
<th>N</th>
<th>S</th>
<th>N</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eve (1;1-2;6)</td>
<td>481</td>
<td>260</td>
<td>76.4</td>
<td>33.0</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sarah (2;0-3;5)</td>
<td>2303</td>
<td>140</td>
<td>70.6</td>
<td>26.2</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adam (2;8-4;14)</td>
<td>3792</td>
<td>760</td>
<td>39.7</td>
<td>32.3</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nina (4;11-5;11)</td>
<td>4042</td>
<td>160</td>
<td>51.5</td>
<td>48.7</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 500</td>
<td>686</td>
<td>243</td>
<td>33.4</td>
<td>31.8</td>
<td>1.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 1000</td>
<td>1800</td>
<td>462</td>
<td>59.3</td>
<td>29.1</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 5000</td>
<td>30000</td>
<td>1400</td>
<td>33.0</td>
<td>24.2</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The theoretical expectations and the empirical measures of overlap agree extremely well... Neither paired t- nor Wilcoxon test reveal significant difference between the two sets of values.*

Yang 2010

Determiner usage case study: D = {a, the} only

*Perhaps a more revealing test is linear regression (Figure 5): a perfect agreement between the two sets of value would have the slope of 1.0, and the actual slope is 1.08 (adjusted $R^2 = 0.9716$). Therefore, we could that the determiner usage data from child language is consistent with the productive rule "DP $\rightarrow$ D N".*

Yang 2010

Determiner usage case study: D = {a, the} only

*Given $N$ unique nouns in a sample of $S$, [a] greater overlap value can be obtained if more nouns occur more than once. That is, words whose probabilities are greater than 1/$S$ can increase the overlap value.*
Zipf's law...allows us to express this cutoff line in terms with
ranks, as the probability of the noun \( n_r \) with rank \( r \) has the
probability of \( 1/(r \cdot \text{HN}) \). The derivation...uses the fact that the
Harmonic Number \( \sum_{i=1}^{N} \frac{1}{i} \) can be approximated by \( \ln N \).

So for Naomi, we expect only the first 2 or 3 ranked nouns to have a non-zero
overlap. For the Brown corpus, we expect only the first 4 or 5 ranked nouns to have a non-zero
overlap.

How do we evaluate the item-based approach, though?
"In the limiting case, the item-based child learner could store the input data in
its entirety and simply retrieve these memorized determiner-noun pairs in
production. Since the input data, which comes from adults, is presumably
productive, children's repetition of it may show the same degree of
productivity."

"Tomasello (2000c, p77) suggests that "...so they simply retrieve that
expression from their stored linguistic experience." Following this line of
reasoning, we consider a learning model that memorizes jointly formed, as
opposed to productively composed, determiner-noun pairs from the input;
presumably these "stored linguistic experience" as such nouns (and
determiners) constitute a large part of adult-child linguistic communication
in every-day life. These pairs will then be sampled directly..."
How do we evaluate the item-based approach, though?

Global memory learner: composite of all children’s input
Local memory learner: drawn just from one particular child’s input

Both sets of overlap values from the two variants of item-based learning... differ significantly from the empirical measures: p < 0.005 for both paired t-test and paired Wilcoxon test. This suggests that children’s use of determiners does not follow the predictions of the item-based learning approach...

Naturally, our evaluation here is tentative since the proper test can be carried out only when the theoretical predictions of item-based learning are made clear. And that is exactly the point: the advocates of item-based learning not only rejected the alternative hypothesis without adequate statistical tests, but also accepted the favored hypothesis without adequate statistical tests.

Case study: Verbal morphology

Few stems appear in a great number of inflections, which, however, never approach anywhere near the maximum number of possible inflections. Moreover, most stems are used very sparsely, the majority of which occur in exactly one inflection.

Example: Spanish verb morphology [1st, 2nd, 3rd person + sg vs. pl]

- Present tense, imperfect aspect, indicative mood
- Past tense, perfect aspect, subjunctive mood
- Present tense, imperfect aspect, indicative mood
- Past tense, perfect aspect, subjunctive mood
- 1s habló habló habló habló
- 2s hablas hablas hablas hablas
- 3s habla habla habla habla
- 1p hablamos hablamos hablamos hablamos
- 2p habláis habláis habláis habláis
- 3p hablan hablan hablan hablan

Survey of inflectional usage data in Italian, Spanish, and Catalan

6 forms = 1st, 2nd, & 3rd person + sg vs. pl

...the logic of the problem remains the same... the diversity of usage depends on the number of opportunities for a verb stem to appear multiple forms, or S/N... children learning Spanish and Catalan show very similar agreement usage to adults—and the S/N ratios are also very similar for these groups.
Case study: Verbal morphology
Survey of inflectional usage data in Italian, Spanish, and Catalan
6 forms = 1st, 2nd, & 3rd person + sg vs. pl

Italian children: 61.8, 11.0, 7.7, 5.5, 3.8, 3.0, 3.0
Spanish children: 60.1, 3.0, 2.9, 3.2, 3.0, 1.8, 2.289
Catalan children: 58.0, 4.1, 7.0, 5.8, 4.6, 4.6, 4.6

Italian adults: 61.8, 11.0, 7.7, 5.5, 3.8, 3.0, 3.0
Spanish adults: 58.0, 4.1, 7.0, 5.8, 4.6, 4.6, 4.6
Catalan adults: 52.5, 7.0, 5.8, 4.6, 4.6, 4.6, 4.6

"Italian children use somewhat more stems in only one form than Italian adults (81.8% vs. 63.9%), but that follows from the S/N ratio (2.544 vs. 1.533). That is, for each verb, the Italian adults have roughly 66% more opportunities to use it than the Italian children, which would account for the discrepancy in the frequency of one-form verbs."

Case study: Verb arguments
"We focus on constructions that involve a transitive verb and its nominal objects, including pronouns and noun phrases. Following the definition of "sentence frame" in Tomasello’s original Verb Island study (1992, p. 242), each unique lexical item in the object position counts as a unique construction for the verb."

Zipfian distribution for top 15 transitive verbs from 1.1 million utterances of child-directed speech...

"...even for large corpora, a verb appears in few constructions frequently and in most constructions infrequently if at all. The observation of Verb Islands, that verbs tend to combine with one or few elements out of a large range, is in fact characteristic of a fully productive verbal syntax system."

Case study: Verb arguments
"For each verb, we count the frequencies of its top 10 most frequent constructions, which are defined as the verb followed a unique lexical item in the object position (e.g., "ask him" and "ask John" are different constructions, following Tomasello 1992)."

Vocabulary: 100 verbs, 100 potential objects [10,000 combinations]
Monte Carlo sampling simulation: ~28,000 samples
Approximate amount of production data: 9.6 million words

Vocabulary: 1500 verbs, 1500 potential objects [2,250,000 combinations]
Monte Carlo sampling simulation: ~1.4 million samples
Approximate amount of production data: ~4.8 billion words (46 years of non-stop talking)

Basic point:
Unlikely to ever see anything except verb islands in production data
Take home points
“For any type of linguistic expression that involve open class items—and that means every type of linguistic expression—modest measures of usage diversity requires extremely large samples.”

“Zipf’s law hints at the inherent limitations in approaches that stress the storage of construction-specific rules or processes…the Zipfian distribution of linguistic combinations…ensure that most “pairings of form and function” simply will never be heard, never mind stored, and those that do appear may do so with sufficiently low frequency such that no reliable storage and use is possible.”

Take home points
“The sparse data problem strikes…and the role of memory in language learning should not be overestimated. In linguistics and cognitive science, of course, the learner’s Zipfian challenge bears another name: the argument from the poverty of stimulus…To attain full linguistic competence, the child learner must overcome the Zipfian distribution and draw generalizations about language on the basis of few and narrow types of linguistic expressions.”