week 2: computational aspects of grammar acquisition

1. P. Adriaans and M. van Zaanen,
   *Computational Grammar Induction for Linguists*,

2. A. Roberts and E. Atwell,
   *Unsupervised grammar inference systems for natural language*,
“In structuralist linguistics, the idea was that you did it procedurally. That plainly didn't work. You get the phones and the phonemes and the morphemes; and it just gets nowhere.”

Mouton de Gruyter (2004)
The structure of the sentence

The boy who the father explained the answer to was honest

according to Chomsky

according to Kayne
implications

Poverty of the Stimulus
- there is not enough data for any learning

Universals
- specific to language
- absolute

Principles and Parameters
- comprehensive
- learnable
1. Introduction
   a grammar: a (possibly infinite) set of sentences with a finite *structural description*.

2. A bird's eye view of CGI
   Gold's Theorem; a few others

3. An initial empirical model

4. Conclusion
   “[...] The time for an integrative effort concerning language acquisition seems right. A cross-fertilization between linguistics, theory of computation and cognitive neuroscience might lead to breakthroughs in one or more of these fields.
the Chomsky hierarchy and natural languages

certain kinds of recursion: 
\{a^n b^n\}

reduplication: 
\{ww|w in Sigma^*\}

multiple agreement: 
\{a^n b^n c^n\}

crossed agreement: 
\{a^n b^m c^n d^m\}
beyond regular – recursion

“counting”:

\[ X \rightarrow aXb \]

\[ X \rightarrow \{\} \]

which generate the strings:

\{\}, ab, aabb, aaabbb, aaaaabbbb, \ldots

(7) if \( S_1 \) then \( S_2 \).
(8) if if \( S_1 \) then \( S_2 \) then \( S_3 \).
(9) if if if \( S_1 \) then \( S_2 \) then \( S_3 \) then \( S_4 \).
beyond regular – recursion

center embedding ("mirror"):

\[ X \rightarrow aXa \]

\[ X \rightarrow bXb \]

\[ X \rightarrow \{ \} \]

which generate the strings:

\{ \}, aa, bb, abba, baab, aaaa, bbbb, aabbaa, abbbba, \ldots

(1) The mouse that the cat that the dog chased bit ran away.
beyond context-free – recursion

“identity”:

\[ S \rightarrow W_iW_i \]

\[ W \rightarrow X \]

\[ X \rightarrow aX \]

\[ X \rightarrow bX \]

\[ X \rightarrow \{ \} \]

which generates the strings:

\{\}, aa, bb, abab, aaaa, bbbb, aabaab, abbabb, …
beyond context-free – recursion

“identity”: crossed agreement

(10) De lerares heeft de knikkers opgeruimd.
    Literal: The teacher has the marbles collected up.
    Gloss: The teacher collected up the marbles.

(11) Jantje heeft de lerares de knikkers helpen opruimen.
    Literal: Jantje has the teacher the marbles help collect up.
    Gloss: Jantje helped the teacher collect up the marbles.

(12) Aad heeft Jantje de lerares de knikkers laten helpen opruimen.
    Literal: Aad has Jantje the teacher the marbles let help collect up.
    Gloss: Aad let Jantje help the teacher collect up the marbles.
beyond context-free – recursion

“identity”: reduplication

Malay *rumah* "house", *rumah-rumah* "houses". Hawaiian has the important example *wiki-wiki*.

Chinese also uses reduplication: 人 *rén* for "person", 人人 *rénrén* for "everybody". Japanese does it too: 時 *toki* "time", *tokidoki* 時々 "sometimes, from time to time".

Adjective reduplication is common in Standard Mandarin, typically denoting emphasis, less acute degree of the quality described, or an attempt at more indirect speech: *xiaoxiao de* 小小的 (small), *chouchou de* 臭臭的 (smelly).

Noun reduplication is found in the southwestern dialect of Mandarin, which is nearly absent in standard Mandarin (Guoyu). For instance, in Sichuan, *baobao* 包包 (handbag) is used whereas Beijing and Guoyu use *bao’r* 包儿.
beyond context-free – recursion

multiple agreement:

Grandia\textsuperscript{mpl} per multos\textsuperscript{mpl}_{\text{ACC}} tenuantur\textsuperscript{3pl} flumina\textsuperscript{mpl} rivos\textsuperscript{mpl}_{\text{ACC}}. (Ov. Rem. 445)
‘Big rivers dissolve into numerous streams’

Multiple agreement on different (verbal) heads is found in languages such as Swahili where subject agreement (=agreement in noun class) is repeated on every element of the verbal complex:

(12) wa-toto \textbf{wa-li-kuwa} \textbf{wa-me-ki-soma} ki-tabu
\texttt{CL2-children} \texttt{CL2-past-be} \texttt{CL2-perf-CL8-read} \texttt{CL8-book}
‘The children had read the book
(Krifka 1995: 1416)
Grammar Induction: the identification of an infinite structure [with a finite structural description] on the basis of a finite number of examples.

Hume and the Problem of Induction

Gold: identification in the limit
☆ Challenger, Learner agree on a class of languages [grammars]
☆ Challenger provides examples [positive, or both positive and negative]
☆ Learner must converge to correct language [possibly after $\infty$ examples]

Theorem 1 (Gold) A class $G$ of grammars is not learnable from positive information if $G$ contains all finite languages and at least one infinite language.
Gold's proof (a sketch)

**Proof Sketch.** Let $L_1 \subset L_2 \subset \ldots$ be a sequence of finite languages, and set $L_\infty = \cup L_i$. Suppose there were a learning algorithm $A$ which either ineffectively or effectively identifies $\{L : L \text{ finite}\} \cup \{L_\infty\}$ in the limit. Then $A$ must identify any finite language in a finite amount of time. So, we can give $A$ a positive presentation that forces it to make an infinite number of mistakes: first present $A$ with enough examples (with repeats) from $L_1$ to make it guess $L_1$. Then, give it enough examples (with repeats) from $L_2$ to force it to guess $L_2$, and so on. Clearly, all our examples will be in $L_\infty$. ■
Gold's results

<table>
<thead>
<tr>
<th>Class</th>
<th>Complete Information</th>
<th>Positive Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recursively Enumerable Sets</td>
<td>Not learnable</td>
<td>Not learnable</td>
</tr>
<tr>
<td>Decidable Rewriting Systems</td>
<td>Learnable</td>
<td>Not learnable</td>
</tr>
<tr>
<td>Super Finite Set</td>
<td>Learnable</td>
<td>Not learnable</td>
</tr>
<tr>
<td>Finite Sets</td>
<td>Learnable</td>
<td>Learnable</td>
</tr>
<tr>
<td>Finite Class of Finite Sets</td>
<td>Learnable</td>
<td>Learnable</td>
</tr>
</tbody>
</table>
nothing Gold can stay

- Under **representative probability distributions**, the class of learnable languages is considerably larger.

- The **real challenge** for a theory of the learnability of natural languages: understand the universal bias that governs the probability distributions of human communication.

- **Universal Bias vs. Universal Grammar**.
nothing Gold can stay

Nature's first green is gold,
Her hardest hue to hold.
Her early leaf's a flower;
But only so an hour.
Then leaf subsides to leaf.
So Eden sank to grief,
So dawn goes down to day.
Nothing gold can stay.

Robert Frost
probability to the rescue

- A source for an initial estimate of relevant probability distributions: **algorithmic complexity theory**.

**basic insight**: objects that are easier to compute have a higher probability.

can address questions such as:  
“which of two candidate grammars is more probable given the data?”

**relevance for language learning**:  
- to teach a language, select examples that are easy to analyze first.

Horning [1969] proved that **probabilistic context-free grammars**  
can be learned from positive data only.
the three main lines of GI research since the 1970s

- **first direction: recursion-theoretic work** – from Gold to Angluin
  - tell-tale sets
  - membership queries
  - equivalence queries
- **results for CFGs basically negative**
the three main lines of GI research since the 1970s

- **second direction:** "simply try"
  
  Wolff (compression; Solomonoff, Rissanen)
  
  Adriaans (EMILE)
  
  van Zaanen (ABL)
the three main lines of GI research since the 1970s

- **third direction: probabilistic learning**

  PCFGs the only class that is learnable & relevant

  Horning's algorithm not feasible for practical application

  the PAC framework:

**Definition 1** Let $F$ be a concept class, $\delta$ ($0 \leq \delta \leq 1$) a confidence parameter, $\epsilon$ ($0 \leq \epsilon \leq 1$) an error parameter. A concept class $F$ is PAC learnable if for all targets $f \in F$ and all probability distributions $P$ on the sample space $U^*$ the learning algorithm $A$ outputs a concept $g \in F$ such that with probability $(1 - \delta)$ it holds that we have a chance on an error with $P(f \Delta g) \leq \epsilon$, where $f \Delta g = (f - g) \cup (g - f)$. 
the three main lines of GI research since the 1970s

- third direction: probabilistic learning
  - PCFGs the only class that is learnable & relevant
  - Horning's proof is not constructive (no algorithm!)
  - the PAC framework:

    no linguistically interesting classes of languages are known to be distribution-free PAC-learnable.

    the distributions in real corpora are heavy-tailed (Zipf): there is at any stage of the sampling process a considerable probability mass in the set of unseen examples.
Zipf's law

log frequency vs. log rank
PAC learning under *simple* distributions

A distribution is **simple** if it is dominated by a recursively enumerable distribution.
Theorem 2 (Levin)

\[- \log m(x) = - \log P_U(x) + O(1) = K(x) + O(1)\]

Here \(m(x)\) is a universal semi measure, \(^9\) \(P_U(x)\) is the universal a priori probability of a binary string \(x\), \(^10\) and \(K(x)\) is the prefix Kolmogorov complexity of the string \(x\), i.e., the length of the shortest prefix-free program that generates \(x\) on a universal Turing machine. \(^11\)

\(^9\) A recursively enumerable semi-measure \(\mu\) is called universal if it recursively dominates every other enumerable semi-measure \(\mu'\), i.e., \(\mu(x) \geq c\mu'(x)\) for a fixed constant \(c\) independent of \(x\). Levin proved that there is a universal enumerable semi-measure. We fix a universal semi-measure \(m(x)\). The semi-measure \(m(x)\) converges to 0 slower than any positive enumerable function which converges to 0. Of course, \(m(x)\) itself is not recursive.

\(^10\) Defined as: \(P_U(x) = \sum_{U(p) = x} 2^{-|p|}\).

\(^11\) The descriptive complexity of a string \(x\) relative to a Turing machine \(T\) and a binary string \(y\) is defined as the shortest program that gives output \(x\) on input \(y\): \(K_T(x|y) = \min\{|p| : p \in \{0,1\}^*, T(p,y) = x\}\). There is a universal Turing machine \(U\), such that for each Turing machine \(T\) there is a constant \(c_T\), such that for all \(x\) and \(y\), we have \(K_U(x|y) \leq K_T(x|y) + c_T\). This definition is invariant up to a constant with respect to different universal Turing machines. Hence we fix a reference universal Turing machine \(U\), and drop the subscript \(U\) by setting \(K(x|y) = K_U(x|y)\). We define: The Kolmogorov complexity of a binary string \(x\) is \(K(x) = K(x|\epsilon)\).
the coding theorem is important, because...

Li & Vitanyi (1991):

**Theorem 3** A concept class $C$ is learnable under $m(x)$ iff $C$ is also learnable under any arbitrary simple distribution $P(x)$ provided the samples are taken according to $m(x)$.

Adriaans [2001] conjecture:

natural languages are **shallow**, that is, they can be learned from a relatively small set of examples the length of which is logarithmic in the Kolmogorov complexity of the grammar:

**Definition 2** A language $G$ is called shallow if it has a tell-tale set $C \subseteq G$ for which $\forall s \in C(|s| \leq c \log K(G))$
• An expression of a type $T$ is **characteristic** for $T$ if it only appears with contexts of type $T$.

• Similarly, a context of a type $T$ is **characteristic** for $T$ if it only appears with expressions of type $T$.

• Let $G$ be a grammar (context-free or otherwise) of a language $L$. $G$ has **context separability** if each type of $G$ has a characteristic context, and **expression separability** if each type of $G$ has a characteristic expression.

• Natural languages seem to be context- and expression-separable.

• This is nothing but stating that languages can define their own concepts internally (...is a noun, ...is a verb).
A class of languages $C$ is **shallow** if for each language $L$ it is possible to find a context- and expression-separable grammar $G$, and a set of sentences $S$ inducing characteristic contexts and expressions for all the types of $G$, such that the size of $S$ and the length of the sentences of $S$ are logarithmic in the descriptive length of $L$ (relative to $C$).

- Seems to hold for natural languages.
the relevance of shallowness:

Theorem 4 Under $m$, sentences from shallow languages can be generated/sampled in polynomial time.

- If $m$ is a good approximation of co-operative linguistic behavior then the possibility that an arbitrary speaker selects a certain binary string decreases exponentially with the complexity of this string;

- if a language is shallow this increases the probability that the language can be learned efficiently.
where is the field now?

- no integration among the three approaches (formal, empirical, PAC)

- interesting problems:
  - understand the nature of the language classes that are learned by existing Grammar Induction algorithms;
  - integrate PAC and PAC-s learnability approaches with the Gold paradigm.
where is the field now?

- human language acquisition is ill-understood:
  - almost no agreement on general principles,
  - or even the basic categories (nouns, verbs)

- some agreement that the explanation sought should:
  - facilitate efficient on-line interpretation;
  - have adequate descriptive power
  - satisfy certain physical and biological constraints
where is the field now?

- a reasonable **working hypothesis**: human language acquisition
  - is **phased**;
  - is **hybrid**, in the sense that different learning strategies might be invoked in different phases,
  - is **cumulative**, in the sense that structures learned in phase $n$ can be used to **bootstrap** structures in phase $n + 1$. 
on bootstrapping

- an important distinction:
  - open-class words;
  - closed-class words;

- functional categories:

  the main problem of learning a language: find out how it uses functional categories

  open-class words easily bootstrapped; not so functional categories;

  a working hypothesis: learn functional categories on the basis of samples taken from limited sets of open classes.
### Possible Algorithmic Aspects of L1 Acquisition

<table>
<thead>
<tr>
<th>Period</th>
<th>Description</th>
<th>Model</th>
<th>Learning Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9 months</td>
<td>Linking acoustics and events babbling</td>
<td>DFA</td>
<td>evidence-based state merging, prosodic bootstrapping</td>
</tr>
<tr>
<td>9–24 months</td>
<td>Children categorize words into word classes and show evidence of early sensitivity to syntax word classes</td>
<td>complex interaction between deixis and babbling</td>
<td>syntactic and semantic bootstrapping</td>
</tr>
<tr>
<td>2–3.5 years</td>
<td>Language meaning and syntax structure is acquired, emergence of recursive rules</td>
<td>context-free grammar induction as first approximation</td>
<td>Seginer, EMILE, ABL</td>
</tr>
</tbody>
</table>
*Computational Grammar Induction for Linguists*

1. Introduction
   a grammar: a (possibly infinite) set of sentences with a finite *structural description*.

2. A bird's eye view of CGI
   Gold's Theorem; a few others

3. An initial empirical model

4. Conclusion
   “[...] The time for an integrative effort concerning language acquisition seems right. A cross-fertilization between linguistics, theory of computation and cognitive neuroscience might lead to breakthroughs in one or more of these fields.”
Computational Grammar Induction for Linguists

For pointers to several very relevant papers, which explain some of the key concepts that A & vZ mention in their all too brief review, see

http://kybele.psych.cornell.edu/forums/TAU-05/ceilidh.html