Handout #5: Learning

1. Goals and outline of this handout

Goals:
How are OT grammars learned? This handout addresses that question from the perspective of the formal study of language learning. For research from an empirically-based perspective, see the references on p. 232 of ATGtOT (nearly exhaustive through 2001) and check ROA (search for “acquisition”).

Outline:
1. Goals and outline of this handout ................................................................. 1
2. The formal study of language learning ........................................................... 1
3. How Tesar and Smolensky (1998) conceive of the learning problem for OT .... 1
4. Grammar learning in OT ............................................................................... 3
5. Error-driven learning ..................................................................................... 5
6. The subset problem and its solution in OT .................................................. 6
7. Learning underlying representations ............................................................ 8
8. References ..................................................................................................... 9

2. The formal study of language learning

The goal of such work is to establish certain results (often through formal proof) about an idealized version of the acquisition problem. The research questions include:

- What algorithms are needed to learn grammars in a given theory?
- What kinds of data do these algorithms need?
- How efficient are the algorithms? How much data and time do they require? (Scaling problem: if algorithm is slow or inefficient in small test cases, how will it do on real languages?)

Ultimately, this line of research should make connections with research on acquisition based on observations or experiments. Indeed, it already has in OT.


Four entities involved in learning:

- Full structural description of output form. Because they assume PARSE/FILL theory, a full SD of the output is also an exact description of the input and the unfaithful I→O mappings. Under Correspondence Theory, a full structural description will analogously include the output form, the input, and the correspondence relation between them. There are other inaudible (or only indirectly audible) properties of the output as well, such as many aspects of prosodic structure.

Revision number: 145
- Overt form. This is the audible part of the output form — what the learner actually hears.
- The grammar. A ranking of UG constraints.
- The lexicon (underlying representations).

Four activities involved in learning:
- Production-directed parsing: Go from input to full SD of output.
- Robust interpretive parsing: Go from overt form to full SD of output.
- Learning a grammar.
- Learning a lexicon.

The following diagram comes from Tesar and Smolensky (1996)

![Diagram](image)

**Figure 3. Problem decomposition, including lexicon.**

Their research strategy:
- It’s too hard to figure all of this out at once, so we’ll make better progress if we divide the problem up into manageable chunks.
- Focus on problem of learning the grammar (ranking of CON) from full structural descriptions.
- (In correspondence theory terms, a “full structural description” is the same as a candidate; input, output, and a correspondence relation between them.)
- Later work vindicates this divide-and-conquer strategy, since making progress on other aspects of the problem while building on Tesar and Smolensky’s results.
4. Grammar learning in OT

The grammar-learning problem as they characterize it:
- The learner has a universal constraint set CON.
- The learner has a set of accurate, consistent, representative data.
- The data are in the form of full structural descriptions. “Full structural descriptions” include all of the information needed to evaluate faithfulness and markedness constraints. This means that we (temporarily) idealize away from two other problems: (i) How is the lexicon learned? (ii) How does the learner discover inaudible but essential properties of surface structure?

Three key insights:
- Use the theory — candidates, GEN, the current constraint hierarchy — to learn the correct grammar. Avoid special acquisition machinery (like parameter-setting triggers) as much as possible. The best theory of acquisition is minimally different from the theory of adult competence.
- Don’t be confused by factorial typology. Dumb idea: compute all permutations of UG constraints, then search for right grammar in this huge space. Smart idea: home in on target grammar by error-driven correction.
- What linguists do to discover OT grammars (find only crucial interactions) is hard. Don’t try do that! Instead, find some grammar that’s guaranteed to be correct.

Connect the immediately preceding statement with the goal of learning a stratified domination hierarchy (stratified partial ordering) like their (22).

Background to their learning algorithm: The Cancellation/Domination Lemma
Suppose two candidates A and B do not incur identical sets of violation-marks. Then A is more harmonic than B if and only if every mark incurred by A which is not cancelled by a mark of B is dominated by an uncanceled mark of B. (Paraphrased from Prince and Smolensky 2004: 261.)

Mark cancellation:
If two candidates both violate the same constraint, then they each cancel the other’s marks on that constraint up to the smaller number of marks that either candidate receives. In the comparison of these two candidates, those marks are uninformative and hence cancelable. Mark cancellation is not mark elimination and it doesn’t make sense except in pairwise comparisons.

The C/D Lemma and constraint demotion:
- The C/D Lemma simply says that every constraint that favors a loser must be dominated by some constraint that favors the winner.
- Constraint demotion brings that state of affairs into being: it locates the loser-favoring constraints and demotes each of them below the highest-ranking winner-favoring constraint.
- Constraint promotion is problematic because often we don’t know which winner-favoring constraint should be promoted above the loser-favoring constraint. Promotion doesn’t fit the logic of the C/D lemma in the way that demotion does.
Recursive Constraint Demotion (RCD) and comparative tableaux:

- Comparative tableaux (Prince 2002) make RCD much easier to understand and to use.
- Prince describes the RCD algorithm in comparative tableau terms as follows:
  1. Locate all constraint columns lacking L. If there are none, quit.
  2. Place these constraints in a stratum ranked just beneath all previously ranked strata.
  3. Remove from consideration all constraints just ranked (tableau columns), as well as the ranking arguments they satisfy (tableau rows).
  4. Recursive step: Go back to [i], applying it to the remaining rows and columns.

How is this different from Tesar and Smolensky’s description of RCD? What’s the source of these differences?

What happens if the data are inconsistent (e.g., because of variation) or CON is insufficient?

Using a comparative tableau, redo Tesar and Smolensky’s (19) and apply RCD to it. (See Prince (2002) for the correct answer.)

(19) Initial data

<table>
<thead>
<tr>
<th>loser/winner pairs</th>
<th>not-yet-ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FILL\text{\textsc{Nuc}}</td>
</tr>
<tr>
<td>(d) ✓ .\textsc{V.CV.}{\textsc{C}}</td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td>.\textsc{V.CVC.}</td>
</tr>
<tr>
<td>(d) ✓ .\textsc{V.CV.}{\textsc{C}}</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>.\textsc{V.CV.}{\textsc{C}}</td>
</tr>
<tr>
<td>(d) ✓ .\textsc{V.CV.}{\textsc{C}}</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>.\textsc{V.CV.C}\textsc{C}</td>
</tr>
</tbody>
</table>

Example for discussion and analysis:

(1) Axininca Campa data (Payne 1981)

\(/i-\text{n-koma-}i/\quad \text{inj.ko.ma.Ti} \quad \text{‘he will paddle’}\)
\(/\text{no-}n-\text{čh-ik-wai-i}/\quad \text{noñ.čh.}k\text{A.wai.Ti} \quad \text{‘I will continue to cut’}\)

Notation:

\(\text{T} = \text{epenthetic } t\)
\(\text{A} = \text{epenthetic } a\)
(2) Constraints in CON (sufficient for present purposes)
   a. ONSET
   b. DEP
   c. MAX
   d. CODA-COND: violated by any coda except a nasal followed by a homorganic stop or affricate
   e. ALIGN-L(Stem, Word): violated by word-initial deletion or epenthesis.
   f. ALIGN-R(Stem, Syllable): violated when a syllable straddles two morphemes.

(3) Some winner-loser pairs of interest
   iñ.ko.ma.Ti > iñ.ko.ma.i
   iñ.ko.ma.Ti > Tiñ.ko.ma.Ti
   iñ.ko.ma.Ti > ko.ma.Ti
   iñ.ko.ma.Ti > iñ.ko.ma.i
   iñ.ko.ma.Ti > iñ.ko.ma
   noñ.čh.i.kA.wai.Ti > noñ.čh.i.kA.wai.Ti
   noñ.čh.i.kA.wai.Ti > noñ.čh.i.wai.Ti

What constraint rankings do these pairs motivate?

Apply RCD to the problem of learning the phonology of Axininca Campa from the data above.

Complexity of the learning algorithm:
   • If there are $N$ constraints in CON, then in the worst case $N^*(N–1)$ demotion steps will be required.

Explain why.

$N^*(N–1) < N!$ for all $N > 3$. What’s the relevance of this?

Further reading:

5. Error-driven learning

Where do the informative losers come from?

Error-driven constraint demotion (Tesar 1998a):
   a. Pick a winner and apply the current grammar to its associated input (production-directed parsing). If the result is different, call it the loser.
   b. You’ve now got an informative winner-loser pair. Use this in the demotion algorithm.

How can we be so confident that this loser is informative? That is, how can we know that this loser will definitely produce at least one demotion (=further learning), unless of course there’s a problem with CON?
Redo Axininca Campa, but use the error-driven algorithm to find the losers.

6. **The subset problem and its solution in OT**

- How does language learning proceed from positive evidence only?
- If the learner’s nascent grammar produces X when X is prohibited in the target language, positive evidence make the learner realize his/her error and retreat from it.
- So learning must always proceed by gradual expansion prompted by positive evidence.
- Subset Principle: “In the special case where one target language is properly contained within another, the point of this condition is to ensure that the acquisition procedure always guesses a subset language if possible, that is, the smallest language that is also compatible with the positive evidence so far encountered.” (Berwick 1985: 236)

**Phonotactic learning in OT:**
- Phonotactic learning probably precedes most or all morphophonemic learning. Phonotactic learners aren’t dealing with alternations, morphology, paradigms, or non-obvious underlying forms.
- Assumption: Phonotactic learner perceives adult form [X] and takes /X/ to be an input to the grammar. The grammar gives the mapping /X/ \(\rightarrow\) [Y]. Goal of learning is to adjust ranking to produce /X/ \(\rightarrow\) [X].

This requires demotion of markedness constraints. Why?
- At that point, the learner’s grammar will accommodate all of the adult phonotactic possibilities.

The subset problem in phonotactic learning:
- The learner’s grammar should accommodate only the adult phonotactic possibilities.
- Why? Because adults have knowledge of what’s forbidden phonotactically.
- Relatively, in accordance with the overall goals of the research program laid out early in §4, we want learning to proceed only from positive evidence so as to avoid any extragrammatical system for inferring so-called indirect negative evidence (Chomsky 1981: 9).

What this means for OT:
- If at any point the ranking produces a superset language — in phonotactics, this means permitting a structure that’s systematically prohibited in the target language — then we’re lost.
- So pains must be taken to ensure that this never happens.

Earliest approach to this problem:
- In the initial state of the grammar, before any learning has taken place, all markedness constraints dominate all faithfulness constraints.
- Succinctly, the initial state is [M >> F].
- There’s an obvious and strong connection with empirical work on acquisition: children’s early productions are unmarked relative to adult phonology. Children’s phonotactic possibilities are a subset of adults’ phonotactic possibilities.
• Child phonology shows activity by markedness constraints that are not visibly active (or at least richly violated) in the ambient language — e.g., Pater’s (1997) analysis of truncation in child phonology.
• There’s also a connection with work on second-language acquisition — see especially Broselow, Chen, and Wang (1998) and Wang (1995) (Mandarin speakers, who have no final obstruents in their native language, “know” that final voiced obstruents are more marked than voiceless ones.)

The insufficiency of the $[M >> F]$ initial state as a solution to the subset problem:
• Hayes (2004), Ito and Mester (1999), McCarthy (1998), Prince and Tesar (2004), and Smith (2000) show that positing the $[M >> F]$ initial state is insufficient to resolve certain subset problems that arise when there are distributional restrictions or output-output faithfulness effects.
• Why?

A problem in phonotactic learning (Prince and Tesar 2004):
• Korean vs. English: In Korean, $s$ and $\dot{s}$ are in complementary distribution (the latter only before $i$). In English, they’re in contrastive distribution.
• Learners of each language hear $\dot{s}i$. What can they conclude?
• To avoid the subset problem, they should only conclude that $\dot{s}i$ is admissible, not that $\dot{s}$ is generally admissible.

\[(4) \ /\dot{s}i/ \rightarrow [\dot{s}i] \text{ mapping (no ranking)}\]

<table>
<thead>
<tr>
<th>$/\dot{s}i/$</th>
<th>$\dot{s}$</th>
<th>$\ast \dot{s}i$</th>
<th>FAITH</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rightarrow \dot{s}i$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sim \dot{s}i$</td>
<td>L</td>
<td>W</td>
<td>1</td>
</tr>
</tbody>
</table>

Explain the problem.

What does RCD do with this tableau?

Why is it insufficient to posit an M >> F initial state?

Biased constraint demotion (Prince and Tesar 2004):
• The M>>F initial state is really just an artifact of a durable but defeasible bias in favor of high-ranking markedness constraints.
• When more than one constraint favors no losers, by preference rank markedness constraints.
• Only rank faithfulness constraints when there are no markedness constraints to rank.
Explain how this proposal is different from the M>>F initial state and how it addresses the problem with (4).

- This is not enough to solve the general problem, however, and a good deal more needs to be said, for which I urge you to consult their paper and Hayes (2004).

7. Learning underlying representations

The problem:
- So far, we’ve only talked about phonotactic learning, where the discovery of underlying representations is trivial, if the goal of the phonotactic learner is an identity map from adult-identical underlying forms to surface forms.
- In morphophonemic learning, learners must figure out nontrivial underlying representations.
- The choice of underlying representations both influences and is influenced by the grammar.
- One bit of good news: phonotactic restrictions and morphophonemic alternations often converge on the same result. This is the Duplication Problem (handout #1), and OT says that the same grammar is responsible for both phonotactics and morphophonemics.
- So the learner with a firm grasp on phonotactics is in a good position to learn morphophonemics, including underlying representations.

What is underlying representation?
- A (the?) central thesis of generative phonology since Chomsky and Halle (1968) is that, except for suppletion, all morphemes have a unique underlying representation.
- OT adopts this assumption — in fact, it must, since OT adopts the central thesis of generative linguistics that a grammar is a function from underlying to surface representation. OT’s special contribution is a hypothesis about the nature of that grammar/function and not about its existence.
- Schane’s (1974) famous Palauan example in (5) shows what kinds of facts a theory of learning has to deal with.

(5) Palauan Vowel Reduction

<table>
<thead>
<tr>
<th>Underlying Present</th>
<th>Future Participle (conservative)</th>
<th>Future Participle (innovative)</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>/daŋ̂ɔb/</td>
<td>[mɔ-ðaŋ̂ɔb]</td>
<td>[dəŋ̂ɔb-ʔ]</td>
<td>[dəŋ̂ɔb-ʔ]</td>
</tr>
<tr>
<td>/teʔib/</td>
<td>[mɔ-teʔiʔɔb]</td>
<td>[təʔiʔ-ʔ]</td>
<td>[təʔiʔ-ʔ]</td>
</tr>
</tbody>
</table>

Some relevant OT literature:
- I obviously can’t review this work here, so I’ll only talk a bit about Tesar and Prince (2004).

Their example (somewhat simplified):
- Two kinds of alternations are potentially involved: final devoicing and intervocalic voicing.
• Four systems of alternations:
  None: mat~mate and nad~nade
  Language with final devoicing only: mat~mate and nat~nade.
  Language with intervocalic voicing only: mat~made and nad~nade
  Language with both processes: mat~made and nat~nade.
• Learner is presumed to have associated alternants with one another and is trying to figure out appropriate underlying forms and constraint rankings.
• Constraints: *d\_σ, *VtV, IDENT(voice).

Result of phonotactic learning (using BCD):
• Language with no alternations (mat~mate and nad~nade): ID(voice) >> *d\_σ, *VtV.
• Language with final devoicing only (mat~mate and nat~nade): *d\_σ >> ID(voice) >> *VtV.
• Language with intervocalic voicing only (mat~made and nad~nade): *VtV >> ID(voice) >> *d\_σ.
• Language with both processes (mat~made and nat~nade): *d\_σ, *VtV >> ID(voice).

Learning underlying forms:
• “When a feature value in a morpheme alternates, the learner considers each value of the feature in turn, trying an underlying form with each feature value, and seeing which one behaves properly with respect to the phonotactic mapping. If one of the feature values works properly, it is adopted for the underlying form.”

Try this with the example given.

Try it out on Palauan, assuming that phonotactic learning has yielded a grammar in which a constraint against unstressed non-schwa vowels (de Lacy 2002) is ranked above faithfulness.

• “A distinct possibility faced by the algorithm is that neither feature value behaves completely correctly, due to incompleteness of the mapping learned from the phonotactics. In that case, for each feature value, the learner engages in further refinement of the mapping in an effort to fully match the data. The learner then selects the combination of underlying form and (refined) mapping that succeeds in matching the data. In this “back and forth” way, the learner navigates the interdependence between underlying forms and mapping, arriving at a correct grammar.”

8. References

Angluin, Dana (1980) Inductive inference of formal languages from positive data. Information and Control 45, 117-35.


