ACL-05

Psychocomputational Models of Human Language Acquisition

Proceedings of the Workshop

29-30 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Introduction

This meeting of the Psychocomputational Models of Human Language Acquisition (PsychoCompLA 2005) workshop is a follow-up meeting of the first PsychoCompLA workshop help in 2004 in Geneva, Switzerland where it was part of the 20th International Conference on Computational Linguistics (COLING 2004). This year, the workshop was part of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005) held in Ann Arbor, Michigan and shared a joint session with the Ninth Conference on Computational Natural Language Learning (CoNLL-2005).

This workshop brings together scientists whose (at least one) line of investigation is to computationally model the process by which humans acquire various aspects of natural language. Progress in this agenda not only directly informs developmental psycholinguistic and linguistic research but will also have the long term benefit of informing applied computational linguistics in areas that involve the automated acquisition of knowledge from a human or human-computer linguistic environment.

The scientific program consisted of two invited talks, one by Brian MacWhinney and another by Mark Steedman, and 10 paper presentations.

We were especially pleased with the high quality of the submissions and would like to thank the authors for submitting their papers, as well as Mirella Lapata (ACL Workshop Committee Chair), Jason Eisner, Philipp Koehn (Publications Chairs) and Dragomir Radev (Local Arrangements Chair) who were extremely helpful (and patient) on more than one occasion.

Alexander Clark
James Cussens
William Gregory Sakas
Aris Xanthos

London, York, New York, and Lausanne, 2005

Organizers:

William Gregory Sakas (Chair), City University of New York, USA
Alexander Clark, Royal Holloway, University of London, UK
James Cussens, University of York, UK
Aris Xanthos, University of Lausanne, Switzerland

Program Committee:

Robert Berwick, MIT, USA
Antal van den Bosch, Tilburg University, The Netherlands
Ted Briscoe, University of Cambridge, UK
Damir Cavar, Indiana University, USA
Nick Chater, University of Warwick, UK
Stephen Clark, University of Edinburgh, UK
Walter Daelemans, University of Antwerp, Belgium and Tilburg University, The Netherlands
Elan Dresher, University of Toronto, Canada
Jeff Elman, University of California, San Diego, USA
Jerry Feldman, University of California, Berkeley, USA
John Goldsmith, University of Chicago, USA
John Hale, University of Michigan, USA
Mark Johnson, Brown University, USA
Vincenzo Lombardo, Universita di Torino, Italy
Paola Merlo, University of Geneva, Switzerland
Sandeep Prasada, City University of New York, USA
Dan Roth, University of Illinois at Urbana-Champaign, USA
Jenny Saffran, University of Wisconsin, Madison, USA
Ivan Sag, Stanford University, USA
Ed Stabler, University of California, Los Angeles, USA
Mark Steedman, University of Edinburgh, UK
Suzanne Stevenson, University of Toronto, Canada
Patrick Sturt, University of Glasgow, UK
Charles Yang, Yale University, USA

Invited Speakers:

Brian MacWhinney, Carnegie Mellon University, USA
Mark Steedman, University of Edinburgh, UK

Table of Contents

Conference Program

Wednesday, June 29, 2005

9:10–9:30 Opening Remarks

9:00–9:30 *The Input for Syntactic Acquisition: Solutions from Language Change Modeling*
Lisa Pearl

9:30–10:00 *Simulating Language Change in the Presence of Non-Idealized Syntax*
W. Garrett Mitchener

10:30–11:00 Break

11:00–11:30 *Using Morphology and Syntax Together in Unsupervised Learning*
Yu Hu, Irina Matveeva, John GoldSmith and Colin Sprague

11:30–12:00 *Refining the SED Heuristic for Morpheme Discovery: Another Look at Swahili*
Yu Hu, Irina Matveeva, John GoldSmith and Colin Sprague

12:00–12:30 *A Connectionist Model of Language-Scene Interaction*
Marshall R. Mayberry, Matthew W. Crocker and Pia Knoeferle

12:30–2:00 Lunch

2:00–2:30 *A Second Language Acquisition Model Using Example Generalization and Concept Categories*
Ari Rappoport and Vera Sheinman

2:30–3:30 Invited Talk

Item Based Constructions and the Logical Problem
Brian MacWhinney

3:30–4:00 Break

Wednesday, June 29, 2005 (continued)

4:00–4:30 *Statistics vs. UG in Language Acquisition: Does a Bigram Analysis Predict Auxiliary Inversion?*
Xuân-Nga Cao Kam, Iglika Stoyneshka, Lidiya Tornyova, William Gregory Sakas and Janet Dean Fodor

4:30–5:00 *Climbing the path to grammar: a maximum entropy model of subject/object learning*
Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli

5:00–5:30 *The Acquisition and Use of Argument Structure Constructions: A Bayesian Model*
Afra Alishahi and Suzanne Stevenson

Thursday, June 30, 2005

Joint Session with CoNLL

9:00–9:50 Invited talk: Mark Steedman

9:50–10:15 *Steps Toward Deep Lexical Acquisition*
Sourabh Niyogi

10:15–10:40 *Intentional Context in Situated Natural Language Learning*
Michael Fleischman and Deb Roy

Remaining CoNLL sessions

11:10–12:25 Morphology

2:10–2:50 Learning methods and architectures

2:50–3:30 Shared Task: Semantic Role Labeling

4:00-6:00 Shared Task:Semantic Role Labeling (con't)

The Input for Syntactic Acquisition: Solutions from Language Change Modeling

Lisa Pearl
Linguistics Department, University of Maryland
1401 Marie Mount Hall, College Park, MD 20742
llsp@wam.umd.edu

Abstract

Sparse data is a well-known problem for any probabilistic model. However, recent language acquisition proposals suggest that the data children learn from is heavily restricted - children learn only from **unambiguous triggers** (Fodor 1998, Dresher 1999, Lightfoot 1999) and **degree-0 data** (Lightfoot 1991). Surprisingly, we show that these conditions are a necessary feature of an accurate language acquisition model. We test these predictions indirectly by developing a mathematical learning and language change model inspired by Yang's (2003, 2000) insights. Our logic is that, besides accounting for how children acquire the adult grammar so quickly, a viable acquisition proposal must also be able to account for how populations change their grammars over time. The language change we examine is the shift in Old English from a strongly Object-Verb (OV) distribution to a strongly Verb-Object (VO) distribution between 1000 A.D. and 1200 A.D., based on data from the YCOE Corpus (Taylor et al. 2003) and the PPCME2 Corpus (Kroch & Taylor 2000). Grounding our simulated population with these historical data, we demonstrate that these acquisition restrictions seem to be both sufficient and necessary for an Old English population to shift its distribution from strongly OV to strongly VO at the right time.

1 Introduction

Empirically investigating what data children attend to during syntactic acquisition is a difficult task. Traditional experimental methods are not feasible on logistical and ethical grounds – we can't simply lock a group of children in a room for two years, restrict their input to whatever we want, and then see if their syntactic acquisition matches normal patterns. However, when we have a simulated group of language learners who follow a quantified model of individual acquisition, this is exactly what we can do – restrict the input to syntactic acquisition in a very specific way and then observe the results.

The individual acquisition model we use is inspired by Yang's (2003, 2000) model of probabilistic learning for multiple grammars. By using this model in a simulated population of individuals, we provide empirical support for two acquisition proposals that restrict children to only heed data that are **unambiguous triggers** (Dresher 1999, Lightfoot 1999, Fodor 1998) and that appear in **degree-0** clauses (Lightfoot 1991). We use language change as a metric of "correct" acquisition, based on the following idea: if the simulated population that has these restrictions behaves just as the real population historically did, the simulated acquisition process is fairly similar to the real acquisition process. Language change is an excellent yardstick for acquisition proposals for exactly this reason – any theory of acquisition must not only be able to account for how children converge to a close approximation of the adult grammar, but also how they can "misconverge" slightly and allow specific types of grammatical change over time. The nature of this "misconvergence" is key. Children must end up with an Internal Language ("grammar") that is close enough - but not *too* close - to the Observable Language (O-Language) in the population so that change can happen at the right pace.

The language change we use as our metric is the shift in Old English from a strongly Object-Verb (OV) distribution to a strongly Verb-Object (VO) distribution between 1000 and 1200 A.D.

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 1–9,
Ann Arbor, June 2005.

The sharpest part of this shift occurs between 1150 and 1200 A.D., based on data from the YCOE Corpus (Taylor et al. 2003) and the PPCME2 Corpus (Kroch & Taylor 2000). We use this corpus data to estimate the initial OV/VO distribution in the modeled population at 1000 A.D. and to calibrate the modeled population's projected OV/VO distribution between 1000 and 1150 A.D. Then, we demonstrate that the restrictions on acquisition seem both sufficient and surprisingly necessary for the simulated Old English population to shift its distribution to be strongly VO by 1200 A.D – and thus match the historical facts of Old English. In this way, we provide empirical support that we would be hard-pressed to get using traditional methods for these acquisition proposals.

The rest of the paper is laid out as follows: section 2 elaborates on the two acquisition proposals of **unambiguous triggers** and **degree-0 data**; section 3 gives specific implementations of these proposals for Old English; section 4 describes the model used to simulate the population of Old English speakers and how the historical corpus data was used; sections 5 and 6 present the results and conclusion.

2 The Acquisition Proposals

The first proposal is set in a Principles and Parameters framework (Chomsky 1981) where the adult grammar consists of a specific set of parameter values and the process of acquisition is figuring out what those values are. An unambiguous trigger (Fodor 1998, Dresher 1999, Lightfoot 1999) is a piece of data from the O-language that unambiguously signals one parameter value over another for a given parameter. Crucially, an unambiguous trigger for value P_1 of parameter P can be parsed only with value P_1 (and not P_2), no matter what *other* parameter values (A, B, C, …) might also be affecting the O-language form of the data. Because an unambiguous trigger corresponds to exactly one parameter P and thus can alter the value of P only, this proposal would allow children to bypass the Credit Problem noted by Dresher (1999), which is the problem of deciding which parameters to update given a particular piece of input. In addition, unambiguous triggers allow the learner to bypass the combinatoric explosion problem that could occur when trying to set **n** parameters. Instead of having to test out $\mathbf{2^n}$ different grammars on the input in the O-languages, the child's language acquisition mechanism simply tests out the n parameters separately by looking for unambiguous triggers for these n parameters in the input from the O-language. Thus, this proposal aids the process of acquiring the adult grammar quickly and correctly. A potential pitfall of this proposal is data sparseness: the quantity of data that fits this very specific restriction might be very small for a parameter P and the child just might not see enough of it for it to have an effect[1].

The second proposal is that children only heed data in **degree-0** clauses (Lightfoot 1991) when they first begin to set their syntactic parameter values. "Degree" refers to the level of embedding, so a degree-0 clause corresponds to a main clause[2].

```
(1)     Jack thought the giant was easy to fool.
        [--Degree-0-]
                      [---------Degree-1---------]
```

The basis for this proposal is that while local grammatical relationships (such as those in degree-0 clauses) provide a lot of information to the learner, degree-0 data tends to be "messier" grammatically – that is, more grammatical processes seem to apply to degree-0 clauses than to degree-1 clauses. The messier status of this data allows the child to converge to a grammar that is *not* exactly the same as the adult grammar. Thus, this proposal focuses on how to allow small grammatical changes to occur in individuals so that larger changes can happen to the population over time. The cost of combining this proposal with the previous one is that the child is now restricted to learn only from **degree-0 unambiguous triggers**, thereby compounding the potential data sparseness problem that unambiguous triggers already have.

[1] In fact, it may well be necessary to restrict the set of parameters relevant for determining if a trigger is unambiguous to some initial pool in order to get any unambiguous triggers at all. A candidate set for the initial pool of parameters might be derived from a hierarchy of parameters along the lines of the one based on cross-linguistic comparison that is described in Baker (2001, 2005).

[2] The exact domain of a degree-0 clause is defined as the main clause and the front of the embedded clause for theory-internal reasons. For a more detailed description and explanation, see Lightfoot (1991).

3 Old English Change

Allowing language change to occur as it historically did is a mark of "correct" acquisition, especially for change involving syntactic parameters that can only be altered during acquisition - any change that builds up in the population must be due to changes that occur during acquisition. The parameter we use in this work is OV/VO word order and the change is a shift in Old English from a strongly OV distribution between 1000 and 1150 A.D. to a strongly VO distribution at 1200 A.D. A strongly OV distribution has many utterances with OV order (2). A strongly VO distribution as many utterances with VO order (3).

(2) he Gode þancode
he God thanked
'He thanked God'
(*Beowulf,* 625)

(3) þa ahof Paulus up his heafod
then lifted Paul up his head
'Then Paul lifted his head up'
(*Blickling Homilies,* 187.35)

Because change can occur only during acquisition, the data children are heeding in their input during acquisition has a massive effect on the population's linguistic composition over time. In this work, we explore the possibility that the data children are heeding during acquisition are the degree-0 unambiguous triggers. For Old English, the unambiguous triggers have the form of (4a) and (5a). Examples of unambiguous triggers of each kind are in (4b-c) and (5b-c).

(4a) Unambiguous OV Trigger
[**Object Verb/Verb-Marker**]$_{\text{VP}}$

(4b) he$_{\text{Subj}}$ **[hyne$_{\textbf{Obj}}$ gebidde$_{\textbf{VerbFinite}}$**
Subj **Obj Verb**
[mid ennum mode]$_{\text{PP}}$]$_{\textbf{VP}}$
PP
(*Ælfric's Letter to Wulfsige,* 87.107)

(4c) we$_{\text{Subj}}$ sculen$_{\text{VerbFinite}}$ **[[ure yfele +teawes]$_{\textbf{Obj}}$**
Subj Verb **Obj**
forl+aten$_{\textbf{Verb-Marker}}$]$_{\textbf{VP}}$
Verb-Marker
(*Alcuin's De Virtutibus et Vitiis,* 70.52)

(5a) Unambiguous VO Trigger
[**Verb/Verb-Marker Object**]$_{\text{VP}}$

(5b) & [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ [**awec+d$_{\textbf{VerbFinite}}$**
PP Subj **Verb**
deade$_{\textbf{Obj}}$ [to life]$_{\text{PP}}$]$_{\text{VP}}$
Obj PP
(*Saint James,* 30.31)

(5c) þa$_{\textbf{Adv}}$ ahof$_{\textbf{VerbFinite}}$ Paulus$_{\textbf{Subj}}$ [**up$_{\textbf{Verb-Marker}}$**
Adv Verb Subj **Verb-Marker**
[his heafod]$_{\textbf{Obj}}$] $_{\textbf{VP}}$
Obj
(*Blickling Homilies,* 187.35)

The Object is adjacent to either a Verb or a Verb-Marker on the appropriate side – the correct O-language order. In addition to this correct "surface order" in the O-language, an unambiguous trigger must also have an **unambiguous derivation** to produce this surface order. This means that no other combination of parameters with the alternate word order value could produce the observed surface order. For example, a *Subject Verb Object* utterance could be produced more than one way because of the Verb-Second (V2) movement parameter which was also available in Old English (as in modern Dutch and German). With V2 movement, the Verb moves from its "underlying" position to the second position in the sentence.

(6a) V2 Movement Ambiguity
Subject Verb Object t_{Verb}. (OV + V2)
Subject Verb t_{Verb} *Object.* (VO + V2)

(6b) Subject Verb Object example
heo$_{\textbf{Subj}}$ cl+ansa+d$_{\textbf{VerbFinite}}$
Subj Verb
[+ta sawle +t+as r+adendan]$_{\textbf{Object}}$
Obj
(*Alcuin De virtutibus et vitiis*, 83.59)

(6c) *Subject Verb* t_{Subj} ***Object*** $\boldsymbol{t_{Verb}}$.
(parsed with OV + V2)

heo$_{\text{Subj}}$ cl+ansa+d$_{\text{VerbFinite}}$ t_{Subj}
Subj Verb
[[+ta sawle +t+as r+adendan]$_{\textbf{Obj}}$
Obj
$\boldsymbol{t_{VerbFinite}}$]$_{\textbf{VP}}$
trace-Vb

(6d) *Subject Verb* t_{Subj} $\boldsymbol{t_{Verb}}$ ***Object***.
(parsed with VO + V2)

heo$_{\text{Subj}}$ cl+ansa+d$_{\text{VerbFinite}}$ t_{Subj}
Subj Verb
[$\boldsymbol{t_{VerbFinite}}$ **[+ta sawle +t+as r+adendan]**$_{\textbf{Object}}$]$_{\textbf{VP}}$
trace-Vb **Obj**

Because of this, a *Subject Verb Object* utterance can be parsed with either word order (OV or VO) and so cannot unambiguously signal either order. Thus, correct surface order alone does not suffice – only an utterance with the correct surface order and that cannot be generated with the competing word order value is an unambiguous trigger[3].

Because V2 movement (among other kinds of movement) can move the Verb away from the Object, Verb-Markers can be used to determine the original position of the Verb with respect to the Object. Verb-Markers include **particles** (‘*up*’), **non-finite complements** to finite verbs (‘shall…*perform*’), some **closed-class adverbials** (‘*never*’), and **negatives** (‘*not*’) as described in Lightfoot (1991).

The curious fact about Old English Verb-Markers (unlike their modern Dutch and German counterparts) is that they were unreliable – often they moved away from the Object as well, leaving nothing Verb-like adjacent to the Object. This turned utterances which potentially were unambiguous triggers for either OV or VO order into ambiguous utterances which could not help acquisition. We term this “trigger destruction,” and it has the effect of making the distribution of OV and VO utterances that the child uses during acquisition (the distribution in the degree-0 unambiguous triggers) *different* from the distribution of the OV and VO utterances in the population. It is this difference that “biases” children away from the distribution in the population and it is this difference that will cause small grammatical changes to accumulate in the population until the larger change emerges – the shift from being strongly OV to being strongly VO. Thus, the question of what data children heed during acquisition has found a very suitable testing ground in Old English.

4 The Model

4.1 The Acquisition Model & Old English Data

The acquisition model in this work is founded on several ideas previously explored in the acquisition modeling and language change literature. First, grammars with opposing parameter values (such as OV and VO order) compete with each other both during acquisition (Clack & Robert 1993) and within a population over time (Pintzuk 2002, among others). Second, population-level change is the result of a build-up of individual-level “misconvergences” (Niyogi & Berwick, 1997, 1996, 1995). Third, individual linguistic behavior can be represented as a probabilistic distribution of multiple grammars. This is the result of multiple grammars competing during acquisition and still existing *after* acquisition.

Multiple grammars in an individual are instantiated as that individual accessing g grammars with probability p_g each (Yang 2003). In our simulation, there are two grammars (g = 2) – one with the OV/VO order set to OV and one with the OV/VO order set to VO. In a stable system with g=1, g_1 has probability p_{g1} = 1 of being accessed and all unambiguous triggers come from this grammar. In the unstable system for our language change, g=2 and g_1 is accessed with probability p_{g1} while g_2 is accessed with probability $p_{g2} = 1 - p_{g1}$. *Both* grammars leave unambiguous triggers in the input to the child.

If the quantity of unambiguous triggers from each grammar is approximately equal, these quantities will effectively cancel each other out (whatever quantity pulls the child towards OV will be counterbalanced by the quantity of triggers pulling the child towards VO). Therefore, the

[3] We note that this could potentially be very resource-intensive to determine since all other interfering parameter values (such as V2) must be taken into account. Hence, there is need for some restriction of what parameters must be initially considered to determine if an utterance contains an unambiguous trigger for a given parameter.

crucial quantity is *how many more* unambiguous triggers one grammar has than the other, since this is the quantity that will not be cancelled out. This is the *advantage* a grammar has over another in the input. Table 1 shows the advantage in the degree-0 (D0) clauses and degree-1 (D1) clauses that the OV grammar has over the VO grammar in Old English at various points in time, based on the data from the YCOE (Taylor et al. 2003) and PPCME2 (Kroch & Taylor 2000) corpora.

	D0 OV Adv	D1 OV Adv
1000 A.D.	**1.6%**	**11.3%**
1000 – 1150 A.D	**0.2%**	**7.7%**
1200 A.D.	**-0.4%**[4]	**-19.1%**

Table 1. OV grammar's advantage in the input for degree-0 (D0) and degree-1 (D1) clauses at various points in time of Old English.

The corpus data shows a 1.6% advantage for the OV grammar in the D0 clauses at 1000 A.D. – which means that only 16 out of every 1000 sentences in the input are actually doing any work for acquisition (and more specifically, pulling the child towards the OV grammar). The data also show that the D1 advantage is much stronger. However, this does not help our learners for two reasons:

a) Based on samples of modern children's input (4K from CHILDES (MacWhinney & Snow 1985) and 4K from young children's stories (for details on this data, see Pearl (2005)), D1 clauses only make up ~16% of modern English children's input. If we assume that the quantity of D1 input to children is approximately the same no matter what time period they live in[5], then our Old English children also heard D1 data in their input ~16% of the time.

b) Our learners can only use D0 data, anyway.

This leads to two questions for the restrictions imposed by the acquisition proposals - a question of sufficiency and a question of necessity. First, we can simply ask if these restrictions on the data children heed are sufficient to allow the Old English population to shift from OV to VO at the appropriate time. Then, supposing that they are, we can ask if these restrictions are necessary to get the job done – that is, will the population shift correctly even if these restrictions do not hold? We can relax both the restriction to learn only from unambiguous triggers and the restriction to learn only from degree-0 clause data – and then see if the population can still shift to a strongly VO distribution on time.

4.2 The Acquisition Model: Implementation

The acquisition model itself is based around the idea of probabilistic access function of binary parameter values (Bock & Kroch 1989) in an individual. For example, if an individual has a function that accesses the VO order value 30% of the time, the utterances generated by that individual would be VO order 30% of the time and OV order 70% of the time. Note that this is the distribution *before* other parameters such as V2 movement alter the order, so the O-language distribution produced by this speaker is *not* 30-70. However, the O-language distribution will still have some unambiguous OV triggers and some unambiguous VO triggers, so a child hearing data from this speaker will have to deal with the conflicting values. Thus, a child will have a probabilistic access function to account for the OV/VO distribution– and acquisition is the process of setting what the VO access probability is, based on the data heard during the critical period.

The VO access value ranges from 0.0 (all OV access) to 1.0 (all VO access). A value of 0.3, for example, would correspond to accessing VO order 30% of the time. A child begins with this value at 0.5, so there is a 50% chance of accessing either OV or VO order.

Two mechanisms help summarize the data the child has seen so far without using up computing resources: the Noise Filter and a modified Batch Learner Method (Yang 2003). The Noise Filter acts as a buffer that separates "signal" from "noise". An unambiguous trigger from the minority grammar is much more likely to be construed as "noise" than an unambiguous trigger from the majority grammar. An example use is

[4] A negative advantage for OV advantage means the VO grammar has the advantage.

[5] At this point in time, we are unaware of any studies that suggest that the composition of motherese, for example, has altered significantly over time.

below with the VO access value set to 0.3 (closer to pure OV than pure VO):

```
6) Noise Filter Use
probabilistic value of VO access = 0.3
if next unambiguous trigger = VO
    = “noise” with 70% chance and ignored
    = “signal” with 30% chance and heeded
if next unambiguous trigger = OV
    = “noise” with 30% chance and ignored
    = “signal” with 70% chance and heeded
```

The initial value of VO access of 0.5, so there is no bias for either grammar when determining what is “noise” and what is “signal”. The modified Batch Learner method deals with how many unambiguous triggers it takes to alter the child’s current VO access value. The more a grammar is in the majority, the smaller the “batch” of its triggers has to be to alter the VO access value (see Table 2). The current VO access value is used to decide whether a grammar is in the majority, and by how much.

VO Value	OV Triggers Required	VO Triggers Required
0.0-0.2	1	5
0.2-0.4	2	4
0.4-0.6	3	3
0.6-0.8	4	2
0.8-1.0	5	1

Table 2. How many unambiguous triggers from each grammar are required, based on what the current VO access value is for the child.

Below is an example of the modified Batch Learner method with the VO access value set to 0.3:

```
7) modified Batch Learner method use
probabilistic value of VO access = 0.3
if next unambiguous trigger = VO
    if 4th VO trigger seen,
        alter value of VO access towards VO
else if next unambiguous trigger = OV
    if 2nd OV trigger seen,
        alter value of VO access towards OV
```

The initial value of 0.5 means that neither grammar requires more triggers than the other at the beginning to alter the current value.

Both mechanisms rely on the probabilistic value of VO access to reflect the distribution of triggers seen so far. The logic is as follows: in order to get to a value below 0.5 (more towards OV), significantly more unambiguous OV triggers must have been seen; in order to get to a value above 0.5 (more towards VO), significantly more unambiguous VO triggers must have been seen.

The individual acquisition algorithm used in the model is below:

```
Initial value of VO access = 0.5
While in critical period
    Get a piece of input from the linguistic environment created by the rest of the population members.
    If input is an unambiguous trigger
        If input passes through Noise Filter
            Increase relevant batch counter
            If counter is at threshold
                Alter current VO access value
```

Note that the final VO access value after the critical period is over does not have to be 0.0 or 1.0 – it may be a value in between. It is supposed to reflect the distribution the child has heard, not necessarily be one of the extreme values.

4.3 Population Model: Implementation

Since individual acquisition drives the linguistic composition of the population, the population algorithm centers around the individual acquisition algorithm:

```
Population age range = 0 to 60
Initial population size = 18000[6]
Initialize members to starting VO access value[7]
At 1000 A.D. and every 2 years until 1200 A.D.
    Members age 59-60 die; the rest age 2 years
    New members age 0 to 1 created
        New members use individual acquisition algorithm to set their VO access value
```

[6] Based on estimates from Koenigsberger & Briggs (1987).

[7] Based on historical corpus data.

4.4 Population Values from Historical Data

We use the historical corpus data to initialize the average VO access value in the population at 1000 A.D., calibrate the model between 1000 and 1150 A.D., and determine how strongly VO the distribution has to be by 1200 A.D. However, note that while the VO access value reflects the OV/VO distribution *before* interference from other parameters causes utterances to become ambiguous, the historical data reflects the distribution *after* this interference has caused utterances to become ambiguous. Table 3 shows how much of the data from the historical corpus is comprised of ambiguous utterances.

Time Period	D0 % Ambig	D1 % Ambig
1000 A.D.	76%	28%
1000-1150 A.D.	80%	25%
1200 A.D.	71%	10%

Table 3. How much of the historical corpus is comprised of ambiguous utterances at various points in time.

We know that either OV or VO order was used to generate all these ambiguous utterances – so our job is to estimate how many of them were generated with the OV order and how many with the VO order. This determines the “underlying” distribution. Once we know this, we can determine what VO access value produced that underlying OV/VO distribution. Following the process detailed in Pearl (2005), we rely on the fact that the D0 distribution is more distorted than the D1 distribution (since the D0 distribution always has more ambiguous triggers). The process itself involves using the difference in distortion between the D0 and D1 distribution to estimate the difference in distortion between the D1 and underlying distribution. Once this is done, we have average VO access values for initialization, calibration, and the target.

Time A.D.	1000	1000-1150	1200
Avg VO	.23	.31	.75

Table 4. Average VO access value in the population at various points in time, based off historical corpus data.

Thus, to satisfy the historical facts, a population must start with an average VO access value of 0.23 at 1000 A.D., reach an average VO access value of 0.31 between 1000 and 1150 A.D., and reach an average VO access value of 0.75 by 1200 A.D.

5 Results

5.1 Sufficient Restrictions

Figure 1 shows the average VO access value over time of an Old English population restricted to learn only from degree-0 unambiguous triggers. These restrictions on acquisition seem sufficient to get the shift from a strongly OV distribution to a strongly VO distribution to occur at the right time. We also note that the sharper population-level change emerges after a build-up of individual-level changes in a growing population.

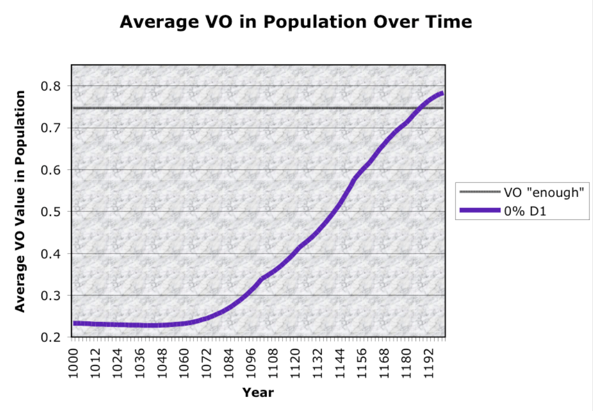


Figure 1. The trajectory of a population restricted to learn only from degree-0 unambiguous triggers.

Thus, we have empirical support for the acquisition proposal since it can satisfy the language change constraints for Old English word order.

5.2 Necessary Restrictions

5.2.1 Unambiguous Triggers

We have shown these restrictions – to learn only from degree-0 unambiguous triggers - are sufficient to get the job done. But are they necessary? We examine the “unambiguous” aspect first – can we still satisfy the language change constraints if we don’t restrict ourselves to unambiguous triggers? This is especially attractive since it may be resource-intensive to determine if an utterance is unambiguous. Instead, we might try simply using surface word order as a trigger. This would create many more triggers in the input - for instance, a *Subject Verb Object* utterance would now be parsed as a VO trigger. Using this definition of trigger, we get the following data from the historical corpus:

	D0 VO Advantage
1000 A.D.	4.8%
1000 – 1150 A.D.	5.5%
1200 A.D.	8.5%

Table 5. Advantage for the VO grammar in the degree-0 (D0) clauses at various times, based on data from the historical corpus.

The most salient problem with this is that even at the earliest point in time when the population is supposed to have a strongly OV distribution, it is the VO grammar – and *not* the OV grammar – that has a significant advantage in the degree-0 data. A population learning from this data would be hard-pressed to remain OV at 1000 A.D., let alone between 1000 and 1150 A.D. Thus, this definition of trigger will not support the historical facts – we *must* keep the proposal which requires unambiguous triggers.

5.2.2 Degree-0 Data

We turn now to the degree-0 data restriction. Recall that the degree-1 data has a much higher OV advantage before 1150 A.D. (see Table 1). It's possible that if children heard enough degree-1 data, the population as a whole would remain OV too long and be unable to shift to a VO "enough" distribution by 1200 A.D. However, the average amount of degree-1 data available to children is about 16% of the input, based on estimates from modern English children's input. Is this small amount enough to keep the Old English population OV too long? With our quantified model, we can determine if 16% degree-1 data causes our population to not be VO "enough" by 1200 A.D. Moreover, we can estimate what the threshold of permissible degree-1 data is so that the modeled Old English population can match the historical facts. Figure 2 displays the average VO access value in 5 Old English populations exposed to different amounts of degree-1 data during acquisition. As we can see, the population with 16% of the input comprised of degree-1 data is *not* able to match the historical facts and be VO "enough" by 1200 A.D. Only populations with 11% or less degree-1 data in the input can.

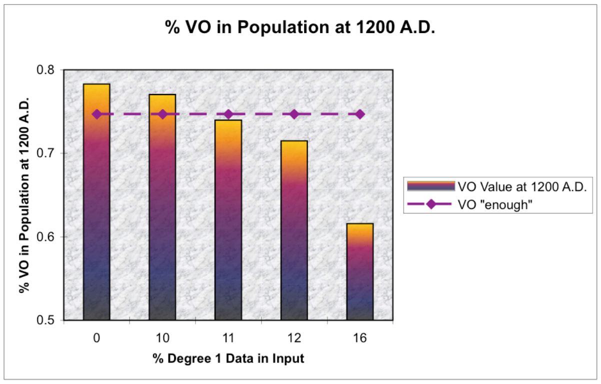


Figure 2: Average VO access values at 1200 A.D. for populations with differing amount of degree-1 data available during acquisition.

This data supports the necessity of the degree-0 restriction since the amount of degree-1 data children hear on average during acquisition (~16%) is too much to allow the Old English population to shift at the right time.

6 Conclusions

Using a probabilistic model of individual acquisition to model a population's language change, we demonstrate the sufficiency and necessity of certain restrictions on individual acquisition. In this way, we provide empirical support for a proposal about what data children are learning from for syntactic acquisition – the degree-0 unambiguous triggers.

Future work will refine the individual acquisition model to explore the connection between the length of the critical period and the parameter in question, including more sophisticated techniques of Bayesian modeling (to replace the current mechanisms of Noise Filter and Batch Learner), and investigate what parameters must be considered to determine if a trigger is "unambiguous". As well, we hope to test the degree-0 unambiguous trigger restriction for other parameters with documented language change, such as the loss of V2 movement in Middle English (Yang 2003, Lightfoot 1999, among others). This type of language change modeling may also be useful for testing proposals about what the crucial data is for phonological acquisition.

Acknowledgement

I am immensely grateful to Amy Weinberg, Charles Yang, Garrett Mitchener, David Lightfoot, Norbert Hornstein, Stephen Crain, Rosalind Thornton, Tony Kroch, Beatrice Santorini, Ann Taylor, Susan Pintzuk, Philip Resnik, Cedric Boeckx, Partha Niyogi, Michelle Hugue, and the audiences of the 28th PLC, DIGS VIII, ICEHL 13, the Cognitive Neuroscience of Language Lunch Talks, and the 2003 Maryland Student Conference. Their good questions, unwavering encouragement, and sensible advice were invaluable. That I may have ignored some of it is my own fault. This work is supported by an NSF Graduate Fellowship.

References

Baker, M. (2001). *The Atoms of Language: The Mind's Hidden Rules of Grammar*. New York, NY: Basic Books.

Baker, M. (2005). Mapping the Terrain of Language Learning. *Language Learning and Development*, 1: 93-129.

Bock, J. & A. Kroch. 1989. The Isolability of Syntactic Processing. In *Linguistic Structure in Language Processing* . Edited by G. Carlson and M. Tannenhaus. Boston: Kluwer.

Chomsky, Noam. (1981). *Lectures on Government and Binding Theory*. Dordrecht: Foris.

Clark, Robin & Ian Roberts (1993). A computational model of language learnability and language change. *Linguistic Inquiry* 24: 299-345.

Dresher, Elan. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30:27-67

Fodor, Janet D. (1998). Unambiguous Triggers. *Linguistic Inquiry,* 29:1-36.

Gibson, Edward & Wexler, Kenneth. (1994). Triggers. *Linguistic Inquiry, 25,* 407-454.

Koenigsberger, H.G. & Briggs, A. (1987). *Medieval Europe, 400-1500.* Longman: New York.

Kroch, Anthony and Taylor, Ann. (2000). The Penn-Helsinki parsed corpus of Middle English. Philadelphia: Department of Linguistics, University of Pennsylvania, 2nd edn. Accessible via http://www.ling.upenn.edu/mideng

Lightfoot, David. (1991). *How to set parameters*. Cambridge, MA: MIT Press.

Lightfoot, David. (1999). *The Development of Language: Acquisition, Change, and Evolution.* Oxford: Blackwell.

MacWhinney, B. & C. Snow. (1985). The Child Language Data Exchange System. *Journal of Child Language* 12: 271-96.

Niyogi, Partha & Berwick, Robert C. (1995). *The logical problem of language change.* AI-Memo 1516, Artificial Intelligence Laboratory, MIT.

Niyogi, Partha & Berwick, Robert C. (1996). A language learning model for finite parameter spaces. *Cognition, 61,* 161-193.

Niyogi, Partha & Berwick, Robert C. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy, 20,* 697-719.

Pearl, Lisa. (2005). The Input to Syntactic Acquisition: Answer from Language Change Modeling. Unpublished Manuscript, University of Maryland, College Park. http://www.wam.umd.edu/~llsp/papers/InputSynAcq.pdf

Pintzuk, Susan (2002). Verb-Object Order in Old English: Variation as Grammatical Competition. *Syntactic Effects of Morphological Change.* Oxford: Oxford University Press.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). The York-Toronto-Helsinki parsed corpus of Old English. York, UK: Department of Language and Linguistic Science, University of York. Available through the Oxford Text Archive.

Yang, Charles. (2000). Internal and external forces in language change. *Language Variation and Change,* 12: 231-250.

Yang, Charles. (2003). *Knowledge and Learning in Natural Language.* Oxford: Oxford University Press.

Simulating Language Change in the Presence of Non-Idealized Syntax

W. Garrett Mitchener
Mathematics Department
Duke University
Box 90320
Durham, NC 27708
wgm@math.duke.edu

Abstract

Both Middle English and Old French had a syntactic property called *verb-second* or *V2* that disappeared. In this paper describes a simulation being developed to shed light on the question of why V2 is stable in some languages, but not others. The simulation, based on a Markov chain, uses *fuzzy grammars* where speakers can use an arbitrary mixture of idealized grammars. Thus, it can mimic the variable syntax observed in Middle English manuscripts. The simulation supports the hypotheses that children use the topic of a sentence for word order acquisition, that acquisition takes into account the ambiguity of grammatical information available from sample sentences, and that speakers prefer to speak with more regularity than they observe in the primary linguistic data.

1 Introduction

The paradox of language change is that on the one hand, children seem to learn the language of their parents very robustly, and yet for example, the English spoken in 800 AD is foreign to speakers of Modern English, and Latin somehow diverged into numerous mutually foreign languages. A number of models and simulations have been studied using historical linguistics and acquisition studies to build on one another (Yang, 2002; Lightfoot, 1999; Niyogi and Berwick, 1996). This paper describes the initial stages of a long term project undertaken in consultation with Anthony Kroch, designed to integrate knowledge from these and other areas of linguistics into a mathematical model of the entire history of English. As a first step, this paper examines the verb-second phenomenon, which has caused some difficulty in other simulations. The history of English and other languages requires simulated populations to have certain long-term behaviors. Assuming that syntax can change without a non-syntactic driving force, these requirements place informative restrictions on the acquisition algorithm. Specifically, the behavior of this simulation suggests that children are aware of the topic of a sentence and use it during acquisition, that children take into account whether or not a sentence can be parsed by multiple hypothetical grammars, and that speakers are aware of variety in their linguistic environment but do not make as much use of it individually.

As discussed in (Yang, 2002) and (Kroch, 1989), both Middle English and Old French had a syntactic rule, typical of Germanic languages, known as *verb-second* or *V2,* in which top-level sentences are re-organized: The finite verb moves to the front, and the topic moves in front of that. These two languages both lost V2 word order. Yang (2002) also states that other Romance languages once had V2 and lost it. However, Middle English is the only Germanic language to have lost V2.

A current hypothesis for how V2 is acquired supposes that children listen for *cue sentences* that cannot be parsed without V2 (Lightfoot, 1999). Specifically, sentences with an initial non-subject topic and

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 10–19,
Ann Arbor, June 2005.

finite verb are the cues for V2:

(1) $[_{CP}$ TopicXP $_{C}$V $[_{IP}$ Subject . . .]]

(2) [[On þis gær] *wolde* [þe king Stephne tæcen. . .]]
[[in this year] *wanted* [the king Stephen seize. . .]]
'During this year king Stephen wanted to seize. . . '
(Fischer et al., 2000, p. 130)

This hypothesis suggests that the loss of V2 can be attributed to a decline in cue sentences in speech. Once the change is actuated, feedback from the learning process propels it to completion.

Several questions immediately arise: Can the initial decline happen spontaneously, as a consequence of purely linguistic factors? Specifically, can a purely syntactic force cause the decline of cue sentences, or must it be driven by a phonological or morphological change? Alternatively, given the robustness of child language acquisition, must the initial decline be due to an external event, such as contact or social upheaval? Finally, why did Middle English and Old French lose V2, but not German, Yiddish, or Icelandic? And what can all of this say about the acquisition process?

Yang and Kroch suggest the following hypothesis concerning why some V2 languages, but not all, are unstable. Middle English (specifically, the southern dialects) and Old French had particular features that obscured the evidence for V2 present in the primary linguistic data available for children:

- Both had underlying subject-verb-object (SVO) word order. For a declarative sentence with topicalized subject, an SVO+V2 grammar generates the same surface word order as an SVO grammar without V2. Hence, such sentences are uninformative as to whether children should use V2 or not. According to estimates quoted in (Yang, 2002) and (Lightfoot, 1999), about 70% of sentences in modern V2 languages fall into this category.

- Both allowed sentence-initial adjuncts, which came before the fronted topic and verb.

- Subject pronouns were different from full NP subjects in both languages. In Middle English, subject pronouns had clitic-like properties that caused them to appear to the left of the finite verb, thereby placing the verb in third position. Old French was a pro-drop language, so subject pronouns could be omitted, leaving the verb first.

The Middle English was even more complex due to its regional dialects. The northern dialect was heavily influenced by Scandinavian invaders: Sentence-initial adjuncts were not used, and subject pronouns were treated the same as full NP subjects.

Other Germanic languages have some of these factors, but not all. For example, Icelandic has underlying SVO order but does not allow additional adjuncts. It is therefore reasonable to suppose that these confounds increase the probability that natural variation or an external influence might disturb the occurrence rate of cue sentences enough to actuate the loss of V2.

An additional complication, exposed by manuscript data, is that the population seems to progress as a whole. There is no indication that some speakers use a V2 grammar exclusively and the rest never use V2, with the decline in V2 coming from a reduction in the number of exclusively V2 speakers. Instead, manuscripts show highly variable rates of use of unambiguously V2 sentences, suggesting that all individuals used V2 at varying rates, and that the overall rate decreased from generation to generation. Furthermore, children seem to use mixtures of adult grammars during acquisition (Yang, 2002). These features suggest that modeling only idealized adult speech may not be sufficient; rather, the mixed speech of children and adults in a transitional environment is crucial to formulating a model that can be compared to acquisition and manuscript data.

A number of models and simulations of language learning and change have been formulated (Niyogi and Berwick, 1996; Niyogi and Berwick, 1997; Briscoe, 2000; Gibson and Wexler, 1994; Mitchener, 2003; Mitchener and Nowak, 2003; Mitchener and Nowak, 2004; Komarova et al., 2001) based on the simplifying assumption that speakers use one grammar exclusively. Frequently, V2 can never be lost in such simulations, perhaps because the learning algorithm is highly sensitive to noise. For example, a simple batch learner that accumulates sample sentences and tries to pick a grammar consistent with all of them might end up with a V2 grammar on the basis of a single cue sentence.

The present work is concerned with developing an improved simulation framework for investigating syntactic change. The simulated population consists of individual simulated people called *agents* that can use arbitrary mixtures of idealized grammars called *fuzzy grammars*. Fuzzy grammars enable the simulation to replicate smooth, population-wide transitions from one dominant idealized grammar to another. Fuzzy grammars require a more sophisticated learning algorithm than would be required for an agent to acquire a single idealized grammar: Agents must acquire usage rates for the different idealized grammars rather than a small set of discrete parameter values.

2 Linguistic specifics of the simulation

The change of interest is the loss of V2 in Middle English and Old French, in particular why V2 was unstable in these languages but not in others. Therefore, the idealized grammars allowed in this simulation will be limited to four: All have underlying subject-verb-object word order, and allow sentence-initial adjuncts. The options are V2 or not, and pro-drop or not. Thus, a grammar is specified by a pair of binary parameter values. For simplicity, the pro-drop parameter as in Old French is used rather than trying to model the clitic status of Middle English subject pronouns.

Sentences are limited to a few basic types of declarative statements, following the degree-0 learning hypothesis (Lightfoot, 1999): The sentence may or may not begin with an adjunct, the subject may be either a full noun phrase or a pronoun, and the verb may optionally require an object or a subject. A verb, such as *rain*, that does not require a subject is given an expletive pronoun subject if the grammar is not pro-drop. Additionally, either the adjunct, the subject, or the object may be topicalized. For a V2 grammar, the topicalized constituent appears just before the verb; otherwise it is indicated only by spoken emphasis.

A fuzzy grammar consists of a pair of beta distributions with parameters α and β, following the convention from (Gelman et al., 2004) that the density for $\mathrm{Beta}(\alpha, \beta)$ is

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1. \tag{3}$$

Each beta distribution controls one parameter in the idealized grammar.[1] The special case of $\mathrm{Beta}(1,1)$ is the uniform distribution, and two such distributions are used as the initial state for the agent's fuzzy grammar. The density for $\mathrm{Beta}(1+m, 1+n)$ is a bump with peak at $m/(m+n)$ that grows sharper for larger values of m and n. Thus, it incorporates a natural critical period, as each additional data point changes the mean less and less, while allowing for variation in adult grammars as seen in manuscripts.

To produce a sentence, an agent with fuzzy grammar $(\mathrm{Beta}(\alpha_1, \beta_1), \mathrm{Beta}(\alpha_2, \beta_2))$ constructs an idealized grammar from a pair of random parameter settings, each 0 or 1, selected as follows. The agent picks a random number $Q_j \sim \mathrm{Beta}(\alpha_j, \beta_j)$, then sets parameter j to 1 with probability Q_j and 0 with probability $1 - Q_j$. An equivalent and faster operation is to set parameter j to 1 with probability μ_j and 0 with probability $1 - \mu_j$, where $\mu_j = \alpha_j/(\alpha_j + \beta_j)$ is the mean of $\mathrm{Beta}(\alpha_j, \beta_j)$.

To learn from a sentence, an agent first constructs a random idealized grammar as before. If the grammar can parse the sentence, then some of the agent's beta distributions are adjusted to increase the probability that the successful grammar is selected again. If the grammar cannot parse the sentence, then no adjustment is made. To adjust $\mathrm{Beta}(\alpha, \beta)$ to favor 1, the agent increments the first parameter, yielding $\mathrm{Beta}(\alpha + 1, \beta)$. To adjust it to favor 0, the agent increments the second parameter, yielding $\mathrm{Beta}(\alpha, \beta + 1)$.

Within this general framework, many variations are possible. For example, the initial state of an agent, the choice of which beta distributions to update for particular sentences, and the social structure (who speaks to who) may all be varied.

[1] The beta distribution is the conjugate prior for using Bayesian inference to estimate the probability a biased coin will come up heads: If the prior distribution is $\mathrm{Beta}(\alpha, \beta)$, the posterior after m heads and n tails is $\mathrm{Beta}(\alpha + m, \beta + n)$.

The simulation in (Briscoe, 2002) also makes use of Bayesian learning, but within an algorithm for which learners switch abruptly from one idealized grammar to another as estimated probabilities cross certain thresholds. The smoother algorithm used here is preferable because children do not switch abruptly between grammars (Yang, 2002). Furthermore, this algorithm allows simulations to include children's highly variable speech. Children learning from each other is thought be an important force in certain language changes; for example, a recent change in the Icelandic case system, known as dative sickness, is thought to be spreading through this mechanism.

3 Adaptation for Markov chain analysis

To the learning model outlined so far, we add the following restrictions. The social structure is fixed in a loop: There are n agents, each of which converses with its two neighbors. The parameters α_j and β_j are restricted to be between 1 and N. Thus, the population can be in one of N^{4n} possible states, which is large but finite.

Time is discrete with each time increment representing a single sentence spoken by some agent to a neighbor. The population is represented by a sequence of states $(X_t)_{t \in \mathbf{Z}}$. The population is updated as follows by a transition function $X_{t+1} = \phi(X_t, U_t)$ that is fed the current population state plus a tuple of random numbers U_t. One agent is selected uniformly at random to be the hearer. With probability p_r, that agent dies and is replaced by a baby in an initial state $(\mathrm{Beta}(1,1), \mathrm{Beta}(1,1))$. With probability $1 - p_r$, the agent survives and hears a sentence spoken by a randomly selected neighbor.

Two variations of the learning process are explored here. The first, called LEARN-ALWAYS, serves as a base line: The hearer picks an idealized grammar according to its fuzzy grammar, and tries to parse the sentence. If it succeeds, it updates any one beta distribution selected at random in favor of the parameter that led to a successful parse. If the parse fails, no update is made. This algorithm is similar to Naive Parameter Learning with Batch (Yang, 2002, p. 24), but adapted to learn a fuzzy grammar rather than an idealized grammar, and to update the agent's knowledge of only one syntactic parameter at a time.

The second, called PARAMETER-CRUCIAL, is the same except that the parameter is only updated if it is *crucial* to the parse: The agent tries to parse the sentence with that parameter in the other setting. If the second parse succeeds, then the parameter is not considered crucial and is left unchanged, but if it fails, then the parameter is crucial and the original setting is reinforced. This algorithm builds on LEARN-ALWAYS by restricting learning to sentences that are more or less unambiguous cues for the speaker's setting for one of the syntactic parameters. The theory of cue-based learning assumes that children incorporate particular features into their grammar upon hearing specific sentences that unambiguously require them. This process is thought to be a significant factor in language change (Lightfoot, 1999) as it provides a feedback mechanism: Once a parameter setting begins to decline, cues for it will become less frequent in the population, resulting in further decline in the next generation. A difficulty with the theory of cue-based learning is that it is unclear what exactly "unambiguous" should mean, because realistic language models generally have cases where no single sentence type is unique to a particular grammar or parameter setting (Yang, 2002, p. 34, 39). The definition of a crucial parameter preserves the spirit of cue-based learning while avoiding potential difficulties inherent in the concept of "unambiguous."

These modifications result in a finite-state Markov chain with several useful properties. It is *irreducible*, which means that there is a strictly positive probability of eventually getting from any initial state to any other target state. To see this, observe that there is a tiny but strictly positive probability that in the next several transitions, all the agents will die and the following sentence exchanges will happen just right to bring the population to the target state. This Markov chain is also *aperiodic*, which means that at any time t far enough into the future, there is a strictly positive probability that the chain will have returned to its original state. Aperiodicity is a consequence of irreducibility and the fact that there is a strictly positive probability that the chain does not change states from one time step to the next. That happens when a hearer fails to parse a sentence, for example. An irreducible aperiodic Markov chain always has a *stationary distribution.* This is a probability distribution on its states, normally denoted π, such that the probability that $X_t = x$ converges to $\pi(x)$ as $t \to \infty$ no matter what the initial state X_0 is. Furthermore, the transition function preserves π, which means that if X is distributed according to π, then so is $\phi(X, U)$. The stationary distribution represents the long term behavior of the Markov chain.

Agents have a natural partial ordering $\succeq$ defined by

$$\begin{gathered}(\mathrm{Beta}(\alpha_1, \beta_1), \mathrm{Beta}(\alpha_2, \beta_2)) \\ \succeq (\mathrm{Beta}(\alpha_1', \beta_1'), \mathrm{Beta}(\alpha_2', \beta_2')) \\ \text{if and only if} \\ \alpha_1 \geq \alpha_1', \beta_1 \leq \beta_1', \alpha_2 \geq \alpha_2', \text{ and } \beta_2 \leq \beta_2'. \end{gathered} \tag{4}$$

This ordering means that the left-hand agent is slanted more toward 1 in both parameters. Not all pairs of agent states are comparable, but there are unique maximum and minimum agent states under this partial ordering,

$$\begin{aligned} A_{\max} &= (\mathrm{Beta}(N, 1), \mathrm{Beta}(N, 1)), \\ A_{\min} &= (\mathrm{Beta}(1, N), \mathrm{Beta}(1, N)), \end{aligned}$$

such that all agent states A satisfy $A_{\max} \succeq A \succeq A_{\min}$. Let us consider two population states X and Y and denote the agents in X by A_j and the agents in Y by B_j, where $1 \leq j \leq n$. The population states may also be partially ordered, as we can define $X \succeq Y$ to mean all corresponding agents satisfy $A_j \succeq B_j$. There are also maximum and minimum population states $X_{\max}$ and $X_{\min}$ defined by setting all agent states to $A_{\max}$ and $A_{\min}$, respectively.

A Markov chain is *monotonic* if the set of states has a partial ordering with maximum and minimum elements and a transition function that respects that ordering. There is a perfect sampling algorithm called *monotonic coupling from the past (MCFTP)* that generates samples from the stationary distribution π of a monotonic Markov chain without requiring certain properties of it that are difficult to compute (Propp and Wilson, 1996). The partial ordering $\succeq$ on population states was constructed so that this algorithm could be used. The transition function ϕ mostly respects this partial ordering, that is, if $X \succeq Y$, then with high probability $\phi(X, U) \succeq \phi(Y, U)$. This monotonicity property is why ϕ was defined to change only one agent per time step, and why the learning algorithms change that agent's knowledge of at most one parameter per time step. However, ϕ does not quite respect $\succeq$, because one can construct X, Y, and U such that $X \succeq Y$ but $\phi(X, U)$ and $\phi(Y, U)$ are not comparable. So, MCFTP does not necessarily produce correctly distributed samples. However, it turns out to be a reasonable heuristic, and until further theory can be developed and applied to this problem, it is the best that can be done.

The MCFTP algorithm works as follows. We suppose that $(U_t)_{t \in \mathbf{Z}}$ is a sequence of tuples of random numbers, and that $(X_t)_{t \in \mathbf{Z}}$ is a sequence of random states such that each X_t is distributed according to π and $X_{t+1} = \phi(X_t, U_t)$. We will determine X_0 and return it as the random sample from the distribution π. To determine X_0, we start at time $T < 0$ with a list of all possible states, and compute their futures using ϕ and the sequence of U_t. If ϕ has been chosen properly, many of these paths will converge, and with any luck, at time 0 they will all be in the same state. If this happens, then we have found a time T such that no matter what X_T is, there is only one possible value for X_0, and that random state is distributed according to π as desired. Otherwise, we continue, starting twice as far back at time $2T$, and so on. This procedure is generally impractical if the number of possible states is large. However, if the Markov chain is monotonic, we can take the shortcut of only looking at the two paths starting at $X_{\max}$ and $X_{\min}$ at time T. If these agree at time 0, then all other paths are squeezed in between and must agree as well.

4 Tweaking

Since this simulation is intended to be used to study the loss of V2, certain long term behavior is desirable. Of the four idealized grammars available in this simulation, three ought to be fairly stable, since there are languages of these types that have retained these properties for a long time: SVO (French, English), SVO+V2 (Icelandic), and SVO+pro-drop (Spanish). The fourth, SVO+V2+pro-drop, ought to be unstable and give way to SVO+pro-drop, since it approximates Old French before it changed. In any case, the population ought to spend most of its time in states where most of the agents use one of the four grammars predominantly, and neighboring agents should have similar fuzzy grammars.

In preliminary experiments, the set of possible sentences did not contain expletive subject pronouns, sentence initial adverbs, or any indication of spoken stress. Thus, the simulated SVO language was a subset of all the others, and SVO+pro-drop was a subset of SVO+V2+pro-drop. Consequently, the PARAMETER-CRUCIAL learning algorithm was unable to learn either of these languages because the non-V2 setting was never crucial: Any sentence that could be parsed without V2 could also be parsed with it. In later experiments, the sentences and grammars were modified to include expletive pronouns, thereby ensuring that SVO is not a subset of SVO+pro-drop or SVO+V2+pro-drop. In addition, marks were added to sentences to indicate spoken stress on the topic. In the simulated V2 languages, topics are always fronted, so such stress can only appear on the initial constituent, but in the simulated non-V2 languages it can appear on any constituent. This modification ensures that no language within the simulation is a subset of any of the others.

The addition of spoken stress is theoretically plausible for several reasons. First, the acquisition of word order and case marking requires children to infer the subject and object of sample sentences, meaning that such thematic information is available from context. It is therefore reasonable to assume that the thematic context also allows for inference of the topic. Second, Chinese allows topics to be dropped where permitted by discourse, a feature also observed in the speech of children learning English. These considerations, along with the fact that the simulation works much better with topic markings than without, suggests that spoken emphasis on the topic provides positive evidence that children use to determine that a language is not V2.

It turns out that the maximum value N allowed for α_j and β_j must be rather large. If it is too small, the population tends to converge to a *saturated* state where all the agents are approximately $\hat{A} = (\mathrm{Beta}(N,N), \mathrm{Beta}(N,N))$. This state represents an even mixture of all four grammars and is clearly unrealistic. To see why this happens, imagine a fixed linguistic environment and an isolated agent learning from this environment with no birth-and-death process. This process is a Markov chain with a single absorbing state $\hat{A}$, meaning that once the learner reaches state $\hat{A}$ it cannot change to any other state: Every learning step requires increasing one of the numerical parameters in the agent's state, and if they are all maximal, then no further change can take place. Starting from any initial state, the agent will eventually reach the absorbing state. The number of states for an agent must be finite for practical and theoretical reasons, but by making N very large, the time it takes for an agent to reach $\hat{A}$ becomes far greater than its life span under the birth-and-death process, thereby avoiding the saturation problem. With $p_r = 0.001$, it turns out that 5000 is an appropriate value for N, and effectively no agents come close to saturation.

After some preliminary runs, the LEARN-ALWAYS algorithm seemed to produce extremely incoherent populations with no global or local consensus on a dominant grammar. Furthermore, MCFTP was taking an extremely long time under the PARAMETER-CRUCIAL algorithm. An additional modification was put in place to encourage agents toward using predominantly one grammar. The best results were obtained by modifying the speaking algorithm so that agents prefer to speak more toward an extreme than the linguistic data would indicate. For example, if the data suggests that they should use V2 with a high probability of 0.7, then they use V2 with some higher probability, say, 0.8. If the data suggests a low value, say 0.3, then they use an even lower value, say 0.2. The original algorithm used the mean μ_j of beta distribution $\mathrm{Beta}(\alpha_j, \beta_j)$ as the probability of using 1 for parameter j. The biased speech algorithm uses $f(\mu_j)$ instead, where f is a sigmoid function

$$f(\mu) = \frac{1}{1 + \exp(2k - 4k\mu)} \tag{5}$$

that satisfies $f(1/2) = 1/2$ and $f'(1/2) = k$. The numerical parameter k can be varied to exaggerate the effect. This modification leads to some increase in coherence with the LEARN-ALWAYS algorithm; it has minimal effect on the samples obtained with the PARAMETER-CRUCIAL algorithm, however MCFTP becomes significantly faster.

The biased speech algorithm can be viewed as a smoother form of the thresholding operation used in (Briscoe, 2002), discussed earlier. An alternative interpretation is that the acquisition process may involve biased estimates of the usage frequencies of syntactic constructions. Language acquisition requires children to impose regularity on sample data, leading to creoles and regularization of vocabulary, for instance (Bickerton, 1981; Kirby, 2001). This addition to the simulation is therefore psychologically plausible.

5 Results

In all of the following results, the bound on α_j and β_j is $N = 5000$, the sigmoid slope is $k = 2$, the probability that an agent is replaced when selected is $p_r = 0.001$, and there are 40 agents in the population configured in a loop where each agent talks to its two neighbors. See Figure 1 for a key to the notation used in the figures.

First, let us consider the base line LEARN-ALWAYS algorithm. Typical sample populations, such as the one shown in Figure 2, tend to be globally and locally incoherent, with neighboring agents favoring completely different grammars. The results are even worse without the biased speech algorithm.

A sample run using the PARAMETER-CRUCIAL learning algorithm is shown in Figure 3. This population is quite coherent, with neighbors generally favoring similar grammars, and most speakers using non-V2 languages. Remember that the picture represents the internal data of each agent, and that their speech is biased to be more regular than their experience. There is a region of SVO+V2 spanning the second row, and a region of SVO+pro-drop on the fourth row with some SVO+V2+pro-drop speakers. Another sample dominated by V2 with larger regions of SVO+V2+pro-drop is shown in Figure 4. A third sample dominated by non-pro-drop speakers is shown in Figure 5. The MCFTP algorithm starts with a population of all $A_{\max}$ and one of $A_{\min}$ and returns a sample that is a possible future of both; hence, both V2 and pro-drop may be lost and gained under this simulation.

In addition to sampling from the stationary distribution π of a Markov chain, MCFTP estimates the chain's *mixing time*, which is how large t must be for the distribution of X_t to be ε-close to π (in total variation distance). The mixing time is roughly how long the chain must run before it "forgets" its initial state. Since this Markov chain is not quite monotonic, the following should be considered a heuristic back-of-the-napkin calculation for the order of magnitude of the time it takes for a linguistic environment to forget its initial state. Figures 3 and 4 require 29 and 30 doubling steps in MCFTP, which indicates a mixing time of around 2^{28} steps of the Markov chain. Each agent has a probability p_r of dying and being replaced if it is selected. Therefore, the probability of an agent living to age m is $(1-p_r)^m p_r$, with a mean of $(1-p_r)/p_r$. For $p_r = 0.001$, this gives an average life span of 999 listening interactions. Each agent is selected to listen or be replaced with probability $1/40$, so the average lifespan is approximately $40,000$ steps of the Markov chain, which is between 2^{15} and 2^{16}. Hence, the mixing time is on the order of $2^{28-16} = 4096$ times the lifespan of an individual agent. In real life, taking a lifespan to be 40 years, that corresponds to at least $160,000$ years. Furthermore, this is an underestimate, because true human language is far more complex and should have an even longer mixing time. Thus, this simulation suggests that the linguistic transitions we observe in real life taking place over a few decades are essentially transient behavior.

6 Discussion and conclusion

With reasonable parameter settings, populations in this simulation are able to both gain and lose V2, an improvement over other simulations, including earlier versions of this one, that tend to always converge to SVO+V2+pro-drop. Furthermore, such changes can happen spontaneously, without an externally imposed catastrophe. The simulation does not give reasonable results unless learners can tell which component of a sentence is the topic. Preliminary results suggest that the PARAMETER-CRUCIAL learning algorithm gives more realistic results than the LEARN-ALWAYS algorithm, supporting the hypothesis that much of language acquisition is based on cue sentences that are in some sense unambiguous indicators of the grammar that generates them. Timing properties of the simulation suggest that it takes many generations for a population to effectively forget its original state, suggesting that further research should focus on the simulation's transient behavior rather than on its stationary distribution.

In future research, this simulation will be extended to include other possible grammars, particularly approximations of Middle English and Icelandic. That should be an appropriate level of detail for studying the loss of V2. For studying the rise of V2, the simulation should also include V1 grammars as in Celtic languages, where the finite verb raises but the topic remains in place. According to Kroch (personal communication) V2 is thought to arise from V1 languages rather than directly from SOV or SVO languages, so the learning algorithm should be tuned so that V1 languages are more likely to become V2 than non-V1 languages.

The learning algorithms described here do not include any bias in favor of unmarked grammatical features, a property that is thought to be necessary for the acquisition of subset languages. One could easily add such a bias by starting newborns with non-uniform prior information, such as $\mathrm{Beta}(1, 20)$ for example. It is generally accepted that V2 is marked based on derivational economy.[2] Pro-drop is more complicated, as there is no consensus on which setting is marked.[3] The correct biases are not obvious, and determining them requires further research.

Further extensions will include more complex population structure and literacy, with the goal of eventually comparing the results of the simulation to data from the Pennsylvania Parsed Corpus of Middle English.

References

Derek Bickerton. 1981. *Roots of Language*. Karoma Publishers, Inc., Ann Arbor.

E. J. Briscoe. 2000. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296.

E. J. Briscoe. 2002. Grammatical acquisition and linguistic selection. In E. J. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Olga Fischer, Ans van Kemenade, Willem Koopman, and Wim van der Wurff. 2000. *The Syntax of Early English*. Cambridge University Press.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

E. Gibson and K. Wexler. 1994. Triggers. *Linguistic Inquiry*, 25:407–454.

Simon Kirby. 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Natalia L. Komarova, Partha Niyogi, and Martin A. Nowak. 2001. The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–59.

Anthony Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

David Lightfoot. 1999. *The Development of Language: Acquisition, Changes and Evolution*. Blackwell Publishers.

W. Garrett Mitchener and Martin A. Nowak. 2003. Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93, January.

W. Garrett Mitchener and Martin A. Nowak. 2004. Chaos and language. *Proceedings of the Royal Society of London, Biological Sciences*, 271(1540):701–704, April. DOI 10.1098/rspb.2003.2643.

W. Garrett Mitchener. 2003. Bifurcation analysis of the fully symmetric language dynamical equation. *Journal of Mathematical Biology*, 46:265–285, March.

Partha Niyogi and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition*, 61:161–193.

Partha Niyogi and Robert C. Berwick. 1997. A dynamical systems model for language change. *Complex Systems*, 11:161–204.

James Gary Propp and David Bruce Wilson. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(2):223–252.

Charles D. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford.

[2] Although Hawaiian Creole English and other creoles front topic and *wh*-word rather than leaving them *in situ*, so it is unclear to what degree movement is marked (Bickerton, 1981).

[3] On one hand, English-speaking children go through a period of topic-drop before learning that subject pronouns are obligatory, suggesting some form of pro-drop is the default (Yang, 2002). On the other hand, creoles are thought to represent completely unmarked grammars, and they are generally not pro-drop (Bickerton, 1981).

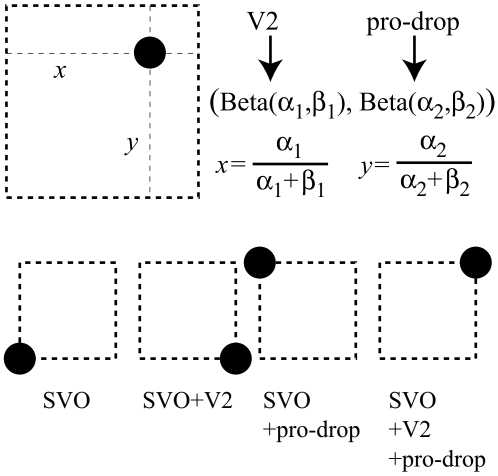

V2
pro-drop
x
y
$\left(\text{Beta}(\alpha_1,\beta_1), \text{Beta}(\alpha_2,\beta_2)\right)$
$x=\frac{\alpha_1}{\alpha_1+\beta_1}$
$y=\frac{\alpha_2}{\alpha_2+\beta_2}$
SVO
SVO+V2
SVO +pro-drop
SVO +V2 +pro-drop

Figure 1: Key to illustrations. Each agent is drawn as a box, with a dot indicating its fuzzy grammar. The means of its beta distributions are used as the coordinates of the dot. The distribution for the V2 parameter is used for the horizontal component, and the distribution for the pro-drop parameter is used for the vertical component. Agents using predominantly one of the four possible idealized grammars have their dot in one of the corners as shown.

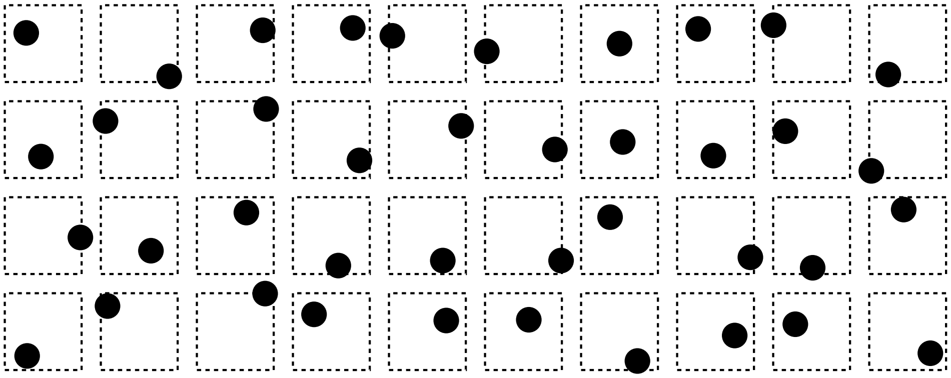


Figure 2: A population of 40 under the LEARN-ALWAYS algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top. The rightmost agent in each row is neighbors with the leftmost agent in the next row up. The bottom left agent is neighbors with the top right agent.

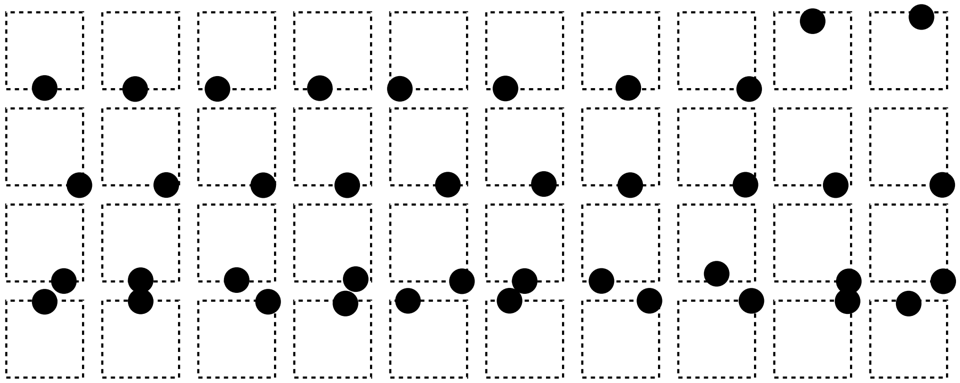


Figure 3: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.

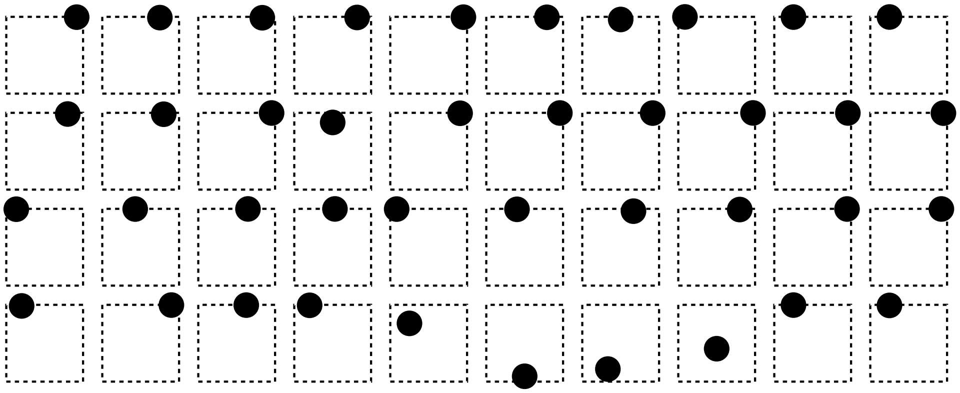


Figure 4: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.

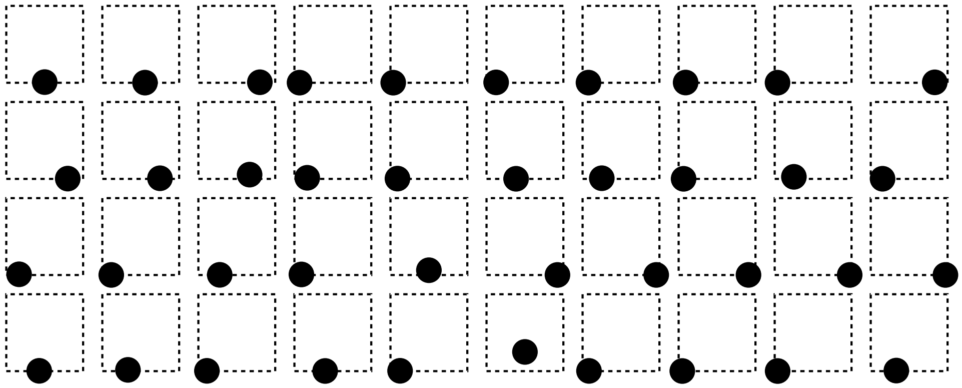


Figure 5: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.

Using Morphology and Syntax Together in Unsupervised Learning

Yu Hu and **Irina Matveeva**
Department of
Computer Science
The University of Chicago
Chicago IL 60637
yuhu@cs.uchicago.edu
matveeva
@uchicago.edu

John Goldsmith
Departments of Linguistics and
Computer Science
The University of Chicago
Chicago IL 60637
ja-goldsmith
@uchicago.edu

Colin Sprague
Department of Linguistics
The University of Chicago
Chicago IL 60637
sprague
@uchicago.edu

Abstract

Unsupervised learning of grammar is a problem that can be important in many areas ranging from text preprocessing for information retrieval and classification to machine translation. We describe an MDL based grammar of a language that contains morphology and lexical categories. We use an unsupervised learner of morphology to bootstrap the acquisition of lexical categories and use these two learning processes iteratively to help and constrain each other. To be able to do so, we need to make our existing morphological analysis less fine grained. We present an algorithm for collapsing morphological classes (signatures) by using syntactic context. Our experiments demonstrate that this collapse preserves the relation between morphology and lexical categories within new signatures, and thereby minimizes the description length of the model.

1 Introduction

Our long term goal is the development of methods which will allow one to produce optimal analyses from arbitrary natural language corpora, where by optimization we understand an MDL (minimum description length; Rissanen, 1989) interpretation of the term: an optimal analysis is one which finds a grammar which simultaneously minimizes grammar length and data compression length. Our specific and primary focus is on *morphology*, and on how knowledge of morphology can be a useful step towards a more complete knowledge of a language's linguistic structure.

Our strategy is based on the following observation: knowing the rightmost suffix of a word is very useful information in inferring (or guessing) a word's part of speech (POS), but due to the ambiguity of many suffixes, it is even better to know both a word's suffix *and* the range of other suffixes that the word's stem appears with elsewhere, i.e., its signature. As we will see below, this conjunction of "better" information is what we call the *signature transform*, and in this paper, we explore how knowledge of signature transform can be merged with knowledge of the context vector to draw conclusions about morphology and syntax.

In the distant future, we would like to be able to use the signature transform in a general process of grammar induction, but that day is not here; we therefore *test* our experiments by seeing how well we are able to predict POS as assigned by an available tagger (TreeTagger; Schmid 1994). In particular, we wish to decrease the uncertainty of a word's POS through the morphological analysis described here. This decrease of uncertainty will enter into our calculation through an *increase* in the probability assigned to our test corpus once the corpus has been augmented with TreeTagger assigned POS tags. But to be clear on our

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 20–27,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

process: we *analyze* a completely raw text morphologically, and use the POS tags from TreeTagger only to evaluate the signature transforms that we generate.

We assume without argument here that any adequate natural language grammar will contain a lexicon which includes both lexical stems which are specified for morphological properties, such as the specific affixes with which they may occur, and affixes associated with lexical categories. We also explicitly note that many affixes are homophonous: they are pronounced (or written) identically, but have different morphological or syntactic characteristics, such as the English plural *–s* and the verbal 3rd person singular present *–s*.

We focus initially on unsupervised learning of *morphology* for three reasons: first, because we already have a quite successful unsupervised morphological learner; second, the final suffix of a word is typically the strongest single indicator of its syntactic category; and third, analysis of a word into a stem T plus suffix F allows us (given our knowledge that the suffix F is a stronger indicator of category than the stem T) to collapse many distinct stems into a single cover symbol for purposes of analysis, simplifying our task, as we shall see.[1] We eschew the use of linguistic resources with hand- (i.e., human-)assigned morphological information in order for this work to contribute, eventually, to a better theoretical understanding of human language acquisition.

We present in this paper an algorithm that modifies the output of the morphology analyzer by combining redundant signatures. Since we ultimately want to use signatures and signature transforms to learn syntactic categories, we developed an algorithm that uses the syntactic contextual information. We evaluate the changes to the morphological analysis from the standpoint of efficient and adequate representation of lexical categories. This paper presents a test conducted on English, and thus can only be considered a preliminary step in the eventually development of a language-independent tool for grammar induction based on morphology. Nonetheless, the concepts that motivate the process are language-independent, and we are optimistic that similar results would be found in tests based on texts from other languages.

In section 2 we discuss the notion of signature and signature transform, and section 3 present a more explicit formulation of the general problem. In section 4 we present our algorithm for signature collapse. Section 5 describes the experiments we ran to test the signature collapsing algorithm, and section 6 presents and discusses our results.

2 Signatures and signature transforms

We employ the unsupervised learning of morphology developed by Goldsmith (Goldsmith, 2001). Regrettably, some of the discussion below depends rather heavily on material presented there, but we attempt to summarize the major points here.

Two critical terms that we employ in this analysis are *signature* and *signature transform*. A *signature* found in a given corpus is a pair of lists: a *stem*-list and a *suffix*-list (or in the appropriate context, a *prefix*-list). By definition of signature σ, the concatenation of every stem in the stem-list of σ with every suffix in the suffix-list of σ is found in the corpus, and a morphological analysis of a corpus can be viewed as a set of signatures that uniquely analyze each word in the corpus. For example, a corpus of English that includes the words *jump, jumps, jumped, jumping, walk, walks, walked,* and *walking* might include the signature σ_1 whose stem list is { *jump, walk* } and whose suffix list is { Ø, ed, ing , s }. For convenience, we label a signature with the concatenation of its suffixes separated by period '.'. On such an analysis, the word *jump* is analyzed as belonging to the signature *Ø.ed.ing.s*, and it bears the suffix Ø. We say, then, that the signature transform of *jump* is *Ø.ed.ing.s_ Ø*, just as the signature transform of *jumping* is *Ø.ed.ing.s_ing*; in general, the signature transform of a word W, when W is morphologically analyzed as stem T followed by suffix F, associated with signature σ, is defined as σ_F.

[1] See Higgins 2002 for a study similar in some ways; Higgins uses morphology as a bootstrap heuristic in one experimental set-up. This paper is heavily indebted to prior work on unsupervised learning of position categories such as Brown et al 1992, Schütze 1997, Higgins 2002, and others cited there.

In many of the experiments described below, we use a corpus in which all words whose frequency rank is greater than 200 have been replaced by their signature transforms. This move is motivated by the observation that high frequency words in natural languages tend to have syntactic distributions poorly predictable by any feature other than their specific identity, whereas the distribution properties of lower frequency words (which we take to be words whose frequency rank is 200 or below) are better predicted by category membership.

In many cases, there is a natural connection between a signature transform and a lexical category. Our ultimate goal is to exploit this in the larger context of grammar induction. For example, consider the signature *Ø.er.ly*, which occurs with stems such as *strong* and *weak*; in fact, words whose signature transform is *Ø.er.ly_ Ø* are adjectives, those whose signature transform is *Ø.er.ly_er* are comparative adjectives, and those whose signature transform is *Ø.er.ly_ly* are adverbs.

The connection is not perfect, however. Consider the signature *Ø.ed.ing.s* and its four signature transforms. While most words whose σ -transform is *Ø.ed.ing.s_s* are verbs (indeed, 3rd person singular present tense verbs, as in *he walks funny*), many are in fact plural nouns (e.g., *walks* in *He permitted four walks in the eighth inning* is a plural noun). We will refer to this problem as the *signature purity problem*–it is essentially the reflex of the ambiguity of suffixes.

In addition, many 3rd person singular present tense verbs are associated with other signature transforms, such as *Ø.ing.s_s*, *Ø.ed.s_s*, and so forth; we will refer to this as the *signature-collapsing problem*, because all other things being equal, we would like to *collapse* certain signatures, such as *Ø.ed.ing.s* and *Ø.ed.ing*, since a stem that is associated with the latter signature *could* have appeared in the corpus with an -s suffix; removing the *Ø.ed.ing* signature and reassigning its stems to the *Ø.ed.ing.s* signature will in general give us a better linguistic analysis of the corpus, one that can be better used in the problem of lexical category induction. This is the reflex of the familiar data sparsity concern.[2]

Since we ultimately want to use signatures and signature transforms to learn syntactic categories, we base the similarity measure between the signatures on the context.

3 A more abstract statement of the problem

A minimum description length (MDL) analysis is especially appropriate for machine learning of linguistic analysis because simultaneously it puts a premium both on analytical simplicity and on goodness of fit between the model and the data (Rissanen 1989).

We will present first the mathematical statement of the MDL model of the morphology, in (1), following the analysis in Goldsmith (2001), followed by a description of the meaning of the terms of the expressions, and then present the modified version which includes additional terms regarding part of speech (POS) information, in (2) and (3).

(1) Morphology

a. Grammar g =

$$\arg\min_{g\in G}\left[Length(g) - \log prob(Data \mid g)\right]$$

b. $Length(g) =$

$$\sum_{t\in T=set\ of\ stems}\left[\log\frac{[W]}{[\sigma(t)]} + \sum_{0\le i<|t|}\log\frac{1}{freq\, t[i]}\right]$$

$$+\sum_{f\in F=set\ of\ affixes}\ \sum_{0\le i<|f|}\log\frac{1}{freq\, f[i]}$$

$$+\sum_{\sigma\in\Sigma}\sum_{f\in\sigma}\left[\log\frac{[\sigma]}{[\sigma\cap f]} + \log\frac{[W]}{[f]}\right]$$

[2] The signature-collapsing problem has another side to it as well. An initial morphological analysis of English will typically give rise to a morphological analysis of words such as *move, moves, moved, moving* with a signature whose stems include *mov* and whose affixes are *e.ed.es.ing*. A successful solution to the signature-collapsing problem will collapse *Ø.ed.ing.s* with *e.ed.es.ing*, noting that *Ø* ~ *e*, *ed ~ed*, *es* ~ *s*, and *ing* ~ *ing* in an obvious sense.

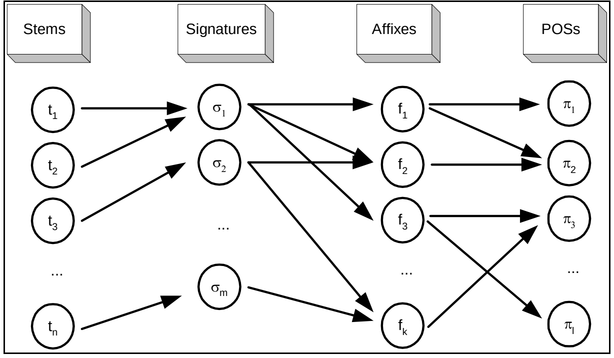
c. $\log prob(Data \mid g) =$

$$\sum_{\substack{w\in Data\\ w=t+f,\sigma}}\left[\begin{array}{l}\log prob(\sigma)\\ +\log prob(t\mid\sigma)\\ +\log prob(f\mid t,\sigma)\end{array}\right]$$

Equation (1a) states that our goal is to find the (morphological) grammar that simultaneously minimizes the sum of its own length and the compressed length of the data it analyzes, while (1b) specifies the grammar length (or model length) as the sum of the lengths of the links between the major components of the morphology: the list of letters (or phonemes) comprising the morphemes, the morphemes (stems and affixes), and the signatures. We use square brackets "[.]" to denote the token counts in a corpus containing a given morpheme or word. The first line of (1b) expresses the notion that each stem consists of a pointer to its signature and a list of pointers to the letters that comprise it; $\sigma(t)$ is the signature associated with stem *t,* and we take its probability to be $\frac{[\sigma(t)]}{[W]}$, the empirical count of the words associated with $\sigma(t)$ divided by the total count of words in the data. The second line expresses the idea that the morphology contains a list of affixes, each of which contains a list of pointers to the letters that comprise it. The third line of (1b) expresses the notion that a signature consists of a list of pointers to the component affixes. (1c) expresses the compressed length of each word in the data.[3]

We now consider extending this model to include part of speech labeling, as sketched in (2). The principal innovation in (2) is the addition of part of speech tags; each affix is associated with one or more POS tags. As we have seen, a path from a particular signature σ to a particular affix f constitutes what we have called a particular signature transform σ_f ; and we condition the probabilities of the POS tags in the data on the preceding signature transformation. As a result, our final model takes the form in (3).

(2)



(3)

a. Grammar g =

$$\arg\min_{g\in G}\left[Length(g)-\log prob(Data\mid g)\right]$$

b. $Length(g) =$

$$\sum_{t\in T=\mathit{set\ of\ stems}}\left[\log\frac{[W]}{[\sigma(t)]}+\sum_{0\le i<|t|}\log\frac{1}{freq\,t[i]}\right]$$

$$+\sum_{f\in F=\mathit{set\ of\ affixes}}\ \sum_{0\le i<|f|}\log\frac{1}{freq\,f[i]}$$

$$+\sum_{\sigma\in\Sigma}\sum_{f\in\sigma}\left[\begin{array}{l}\log\frac{[\sigma]}{[\sigma\cap f]}+\log\frac{[W]}{[f]}+\\ \sum_{\pi\in\Pi}\log\frac{[f\cap\sigma]}{[f\cap\sigma\cap\pi]}\end{array}\right]$$

c. $\log prob(Data \mid g) =$

$$\sum_{\substack{w\in Data\\ w=t+f,\sigma}}\left[\begin{array}{l}\log prob(\sigma)+\log prob(t\mid\sigma)\\ +\log prob(f\mid t,\sigma)\\ +\log prob(\pi\mid\sigma,f)\end{array}\right]$$

The differences between the models are found in the added final term in (3b), which specifies the information required to predict, or specify, the part of speech given the signature

[3] We do not sum over all occurrences of a word in the corpus; we count the compressed length of each word type found in the corpus. This decision was made based on the observation that the (compressed length of the) data term grows much faster than the length of the grammar as the corpus gets large, and the loss in ability of the model to predict word frequencies overwhelms any increase in model simplicity when we count word tokens in the data terms. We recognize the departure from the traditional understanding of MDL here, and assume the responsibility to explain this in a future publication.

transform, and the corresponding term in the corpus compression expression (3c).

The model in (3) implicitly assumes that the true POSs are known; in a more complete model, the POSs play a direct role in assigning a higher probability to the corpus (and hence a smaller compressed size to the data). In the context of such a model, an MDL-based learning device searches for the best assignment of POS tags over all possible assignments. Instead of doing that in this paper, we employ the TreeTagger (Schmid, 1994) based tags (see section 5 below), and make the working assumption that optimization of description length over all signature-analyses and POS tags can be approximated by optimization over all signature-analyses, given the POS tags provided by TreeTagger.

4 The collapsing of signatures

We describe in this section our proposed algorithm, using context vectors to collapse signatures together, composed of a sequence of operations, all but the first of which may be familiar to the reader:

Replacement of words by signature-transforms: The input to our algorithm for collapsing signatures is a modified version of the corpus which integrates the (unsupervised) morphological analyses in the following way. First of all, we leave unchanged the 200 most frequent words (word types). Next, we *replace* words belonging to the K most reliable signatures (where K=50 in these experiments) by their associated *signature transforms*, and we in effect *ignore* all other words, by replacing them by a distinguished "dummy" symbol. In the following, we refer to our high frequency words and signature transforms together as *elements*—so an *element* is any member of the transformed corpus other than the "dummy".

Context vectors based on mutual information: By reading through the corpus, we populate both left and right context vectors for each element (=signature-transform and high-frequency word) by observing the elements that occur adjacent to it. The feature indicating the appearance of a particular word on the *left* is always kept distinct from the feature indicating the appearance of the same word on the *right*. The features in a context vector are thus associated with the members of the element vocabulary (and indeed, each member of the element vocabulary occurs as two features: one on the left, one on the right). We assign the value of each feature y of x's context vector as the pointwise mutual information of the corresponding element pair (x, y), defined as $\log \frac{pr(x,y)}{pr(x)pr(y)}$.

Simplifying context vectors with "idf": In addition, because of the high dimensionality of the context vector and the fact that some features are more representative than others, we trim the original context vector. For each context vector, we sort features by their values, and then keep the top N (in general, we set N to 10) by setting these values to 1, and all others to 0. However, in this resulting *simplified context vector,* not all features do equally good jobs of distinguishing syntactical categories. As Wicentowski (2002) does in a similar context, we assign a weight w_{f_i} to each feature f_i in a fashion parallel to inverse document frequency (*idf;* see Sparck Jones 1973), or $\log \frac{\#total\ distinct\ elements}{\#elements\ this\ feature\ appears\ in}$. We view these as the diagonal elements of a matrix M (that is, $m_{i,i} = w_{f_i}$). We then check the similarity between two simplified context vectors by computing the weighted sum of the dot product of them. That is, given two simplified context vectors *c* and *d*, their *similarity* is defined as c^TMd. If this value is larger than a threshold θ that is set as one parameter, we deem these two context vectors to be similar. Then we determine the similarity between elements by checking whether both left and right simplified context vectors of them are similar (i.e., their weighted dot products exceed a threshold θ). In the experiments we describe below, we explore four settings θ for this threshold: 0.8 (the most "liberal" in allowing greater signature transform collapse, and hence greater signature collapse), 1.0, 1.2, and 1.5.

Calculate signature similarity: To avoid considering many unnecessary pairs of signatures, we narrow the candidates into signature pairs in which the suffixes of one constitute a subset of suffixes of the other, and we set a limit to the permissible difference in the lengths of the signatures in the collapsed pairs, so that the difference in *number* of affixes cannot exceed 2. For each such pair, if all corresponding *signature transforms* are similar in the sense defined in the preceding paragraph, we deem the two *signatures* to be similar.

Signature graph: Finally, we construct a *signature graph*, in which each signature is represented as a vertex, and an edge is drawn between two signatures iff they are similar, as just defined. In this graph, we find a number of cliques, each of which, we believe, indicates a cluster of signatures which should be collapsed. If a signature is a member of two or more cliques, then it is assigned to the largest clique (i.e., the one containing the largest number of signatures).[4]

5 Experiments

We obtain the morphological analysis of the Brown corpus (Kučera and Francis, 1967) using the Linguistica software (http://linguistica.uchicago.edu), and we use the TreeTagger to assign a Penn TreeBank-style part-of-speech tag to each token in the corpus. We then carry out our experiment using the Brown corpus modified in the way we described above. Thus, for each token of the Brown corpus that our morphology analyzer analyzed, we have the following information: its stem, its signature (i.e., the signature to which the stem is assigned), the suffix which the stem attains in this occurrence of the word (hence, the signature-transform), and the POS tag. For example, the token *polymeric* is analyzed into the stem *polymer* and the suffix *ic*, the stem is assigned to the signature *Ø.ic.s*, and thus this particular token has the signature transform *Ø.ic.s_ic*. Furthermore, it was assigned POS-tag *JJ*, so that we have the following entry: "polymeric JJ Ø.ic.s_ic".

Before performing signature collapsing, we calculate the description length of the morphology and the compressed length of the words that our algorithm analyzes and call it *baseline description length* (DL_0).

Now we apply our signature collapsing algorithm under several different parameter settings for the similarity threshold θ, and calculate the description length DL^{θ} of the resulting morphological and lexical analysis using (3). We know that the smaller the set of signatures, the smaller is the cost of the model. However, a signature collapse that combines signatures with different distributions over the lexical categories will result in a high cost of the data term (3c). The goal was therefore to find a method of collapsing signatures such that the reduction in the model cost will be higher than the increase in the compressed length of the data so that the total cost will decrease.

As noted above, we perform this operation iteratively, and refer to the description length of the i^th^ iteration, using a threshold θ, as $DL^{\theta}_{iter=i}$.

We used random collapsing in our experiments to ensure the expected relationship between appropriate collapses and description length. For each signature collapsing, we created a parallel situation in which the *number* of signatures collapsed is the same, but their *choice* is random. We calculate the description length using this "random" analysis as DL^{θ}_{random}. We predict that this random collapsing will not produce an improvement in the total description length.

[4] Our parameters are by design restrictive, so that we declare only few signatures to be similar, and therefore the cliques that we find in the graph are relatively small. One way to enlarge the size of collapsed signatures would be to loosen the similarity criterion. This, however, introduces too many new edges in the signatures graph, leading in turn to spurious collapses of signatures. We take a different approach, and apply our algorithms iteratively. The idea is that if in the first iteration, two cliques did not have enough edges between their elements to become a single new signature, they may be more strongly connected in the second iteration if many of their elements are sufficiently similar. On the other hand, cliques that were dissimilar in the first iteration remain weakly connected in the second.

6 Results and discussion

Table 1 presents the description length, broken into its component terms (see (3)), for the baseline case and the alternative analyses resulting from our algorithm. The table shows the total description length of the model, as well as the individual terms: the signature term DL(σ), the suffix term DL(F), the lexical categories term, DL(P), total morphology, DL(M), and the compressed length of the data, DL(D). We present results for two iterations for four threshold values (θ=0.8,1.0,1.2,1.5) using our collapsing algorithm.

Table 2 presents DL_{random}^{θ} derived from the random collapsing, in a fashion parallel to Table 1. We show the results for only one iteration of random collapsing, since the first iteration already shows a substantial increase in description length.

Figure 1 and Figure 2 present graphically the total description length from Tables 1 and 2 respectively. The reader will see that all collapsing of signatures leads to a shortening of the description length of the morphology *per se*, and an increase in the compressed length of the data. This is an inevitable formal consequence of the MDL-style model used here. The empirical question that we care about is whether the combined description length increases or decreases, and what we find is that when collapsing the signatures in the way that we propose to do, the combined description length decreases, leading us to conclude that this is, overall, a superior linguistic description of the data. On the other hand, when signatures are collapsed randomly, the combined description length increases. This makes sense; randomly decreasing the formal simplicity of the grammatical description should *not* improve the overall analysis. Only an increase in the formal simplicity of a grammar that is grammatically sensible should have this property. Since our goal is to develop an algorithm that is completely data-driven and can operate in an

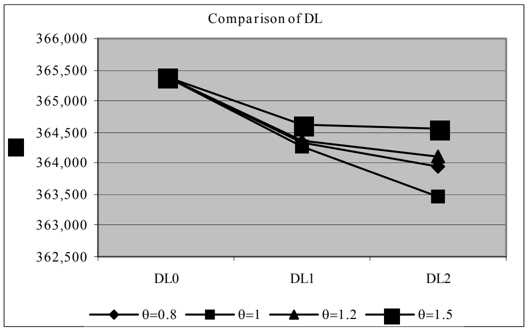


Figure 1 Comparison of DL, 2 iterations and 4 threshold values

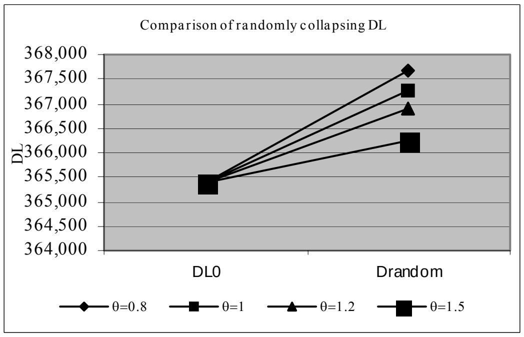


Figure 2 Comparison of DLs with random collapse of signatures (see text)

	DL$_0$	$DL_{iter=1}^{\theta=0.8}$	$DL_{iter=2}^{\theta=0.8}$	$DL_{iter=1}^{\theta=1.0}$	$DL_{iter=2}^{\theta=1.0}$	$DL_{iter=1}^{\theta=1.2}$	$DL_{iter=2}^{\theta=1.2}$	$DL_{iter=1}^{\theta=1.5}$	$DL_{iter=2}^{\theta=1.5}$
#σ	50	41	35	41	34	44	42	46	45
DL(σ)	47,630	45,343	42,939	45,242	43,046	44,897	44,355	46,172	45,780
DL(F)	160	156	156	153	143	158	147	163	164
DL(P)	2,246	2,087	1,968	2,084	1,934	2,158	2,094	2,209	2,182
DL(M)	50,218	47,768	45,244	47,659	45,304	47,395	46,777	48,724	48,306
DL(D)	315,165	316,562	318,687	316,615	318,172	316,971	317,323	315,910	316,251
Total DL	365,383	364,330	363,931	364,275	363,476	364,367	364,101	364,635	364,558

Table 1. DL and its individual components for baseline and the resulting cases when collapsing signatures using our algorithm.

	DL$_0$	$DL_{random}^{\theta=0.8}$	$DL_{random}^{\theta=1.0}$	$DL_{random}^{\theta=1.2}$	$DL_{random}^{\theta=1.5}$
#σ	50	41	41	44	46
DL(σ)	47,630	44,892	45,126	45,788	46,780
DL(F)	160	201	198	187	177
DL(P)	2,246	2,193	2,195	2,212	2,223
DL(M)	50,218	47,468	47,700	48,369	49,362
DL(D)	315,165	320,200	319,551	318,537	316,874
Total DL	365,383	367,669	367,252	366,907	366,237

Table 2. DL and its individual components for baseline and the resulting cases when collapsing signatures randomly.

unsupervised fashion, we take this evidence as supporting the appropriateness of our algorithm as a means of collapsing signatures in a grammatically and empirically reasonable way.

We conclude that the collapsing of signatures on the basis of similarity of context vectors of signature transforms (in a space consisting of high frequency words and signature transforms) provides us with a useful and significant step towards solving the signature collapsing problem. In the context of the broader project, we will be able to use signature transforms as a more effective means for projecting lexical categories in an unsupervised way.

As Table 1 shows, we achieve up to 30% decrease in the number of signatures through our proposed collapse. We are currently exploring ways to increase this value through powers of the adjacency matrix of the signature graph.

In other work in progress, we explore the equally important *signature purity* problem in graph theoretic terms: we *split* ambiguous signature transforms into separate categories when we can determine that the edges connecting left-context features and right-context features can be resolved into two sets (corresponding to the distinct categories of the transform) whose left-features have no (or little) overlap and whose right features have no (or little) overlap. We employ the notion of minimum cut of a weighted graph to detect this situation.

References

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics,* 18(4): 467-479.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics,* 27(2): 153-198.

Higgins, Derrick. 2002. *A Multi-modular Approach to Model Selection in Statistical NLP*. University of Chicago Ph.D. thesis.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees.. *International Conference on New Methods in Language Processing*

Kucera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press.

Rissanen, Jorma. 1989. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Schütze, Hinrich. 1997. *Ambiguity Resolution in Language Learning*. CSLI Publications. Stanford CA.

Sparck Jones, Karen. 1973. Index term weighting. *Information Storage and Retrieval* 9:619-33.

Wicentowski, Richard. 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework.* Johns Hopkins University Ph.D. thesis.

The SED heuristic for morpheme discovery: a look at Swahili

Yu Hu and **Irina Matveeva**
Department of
Computer Science
The University of Chicago
Chicago IL 60637
yuhu@cs.uchicago.edu
matveeva
@uchicago.edu

John Goldsmith
Departments of Linguistics and
Computer Science
The University of Chicago
Chicago IL 60637
ja-goldsmith
@uchicago.edu

Colin Sprague
Department of Linguistics
The University of Chicago
Chicago IL 60637
sprague
@uchicago.edu

Abstract

This paper describes a heuristic for morpheme- and morphology-learning based on string edit distance. Experiments with a 7,000 word corpus of Swahili, a language with a rich morphology, support the effectiveness of this approach.

1 Introduction

This paper describes work on a technique for the unsupervised learning of the morphology of natural languages which employs the familiar string edit distance (*SED*) algorithm (Wagner and Fischer 1974 and elsewhere) in its first stage; we refer to it here as the *SED heuristic*. The heuristic finds 3- and 4-state finite state automata (*FSA*s) from untagged corpora. We focus on Swahili, a Bantu language of East Africa, because of the very high average number of morphemes per word, especially in the verbal system, a system that presents a real challenge to other systems discussed in the literature.[1]

In Section 2, we present the SED heuristic, with precision and recall figures for its application to a corpus of Swahili. In Section 3, we propose three elaborations and extensions of this approach, and in Section 4, we describe and evaluate the results from applying these extensions to the corpus of Swahili.[2]

2 SED-based heuristic

Most systems designed to learn natural language morphology automatically can be viewed as being composed of an initial heuristic component and a subsequent explicit model. The initial or *bootstrapping* heuristic, as the name suggests, is designed to rapidly come up with a set of candidate strings of morphemes, while the model consists of an explicit formulation of either (1) what constitutes an adequate morphology for a set of data, or (2) an objective function that must be optimized, given a corpus of data, in order to find the correct morphological analysis.

The best known and most widely used heuristic is due to Zellig Harris (1955) (see also Harris (1967) and Hafer and Weiss (1974) for an evaluation based on an English corpus), using a notion that Harris called successor frequency (henceforth, *SF*). Harris' notion can be succinctly described in contemporary terms: if we encode all of the data in the data structure known as a trie, with each node in the trie dominating all strings which share a common

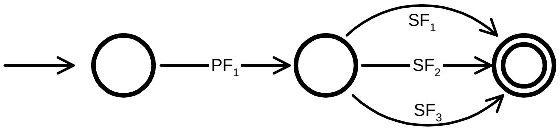
[1] An earlier version of this paper, with a more detailed discussion of the material presented in Section 3, is available at Goldsmith et al (2005).

[2] SED has been used in unsupervised language learning in a number of studies; see, for example, van Zaanen (2000) and references there, where syntactic structure is studied in a similar context. To our knowledge, it has not been used in the context of morpheme detection.

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 28–35,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

string prefix,[3] then each branching node in the trie is associated with a morpheme break. For example, a typical corpus of English may contain the words *governed, governing, government, governor*, and *governs*. If this data is encoded in the usual way in a trie, then a single node will exist in the trie which represents the string prefix *govern* and which dominates five leaves corresponding to these five words. Harris's SF-based heuristic algorithm would propose a morpheme boundary after *govern* on this basis. In contemporary terms, we can interpret Harris's heuristic as providing *sets* of simple finite state automata, as in (1), which generate a string prefix (PF_1) followed by a set of string suffixes (SF_i) based on the measurement of a successor frequency greater than 1 (or some threshold value) at the string position following PF_1.

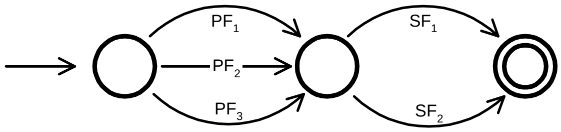
(1)



A variant on the SF-based heuristic, predecessor frequency (henceforth, PF), calls for encoding words in a trie from right to left. In such a PF-trie, each node dominates all strings that share a common string suffix. In general, we expect SF to work best in a suffixing language, and PF to work best in prefixing language; Swahili, like all the Bantu languages, is primarily a prefixing language, but it has a significant number of important suffixes in both the verbal and the nominal systems.

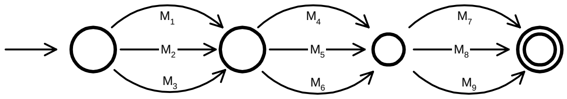
Goldsmith (2001) argues for using the discovery of signatures as the bootstrapping heuristic, where a *signature* is a maximal set of stems and suffixes with the property that all combinations of stems and suffixes are found in the corpus in question. We interpret Goldsmith's signatures as extensions of FSAs as in (1) to FSAs as in (2); (2) characterizes Goldsmith's notion of signature in term of FSAs. In particular, a signature is a set of forms that can be characterized by an FSA of 3 states.

(2)



We propose a simple alternative heuristic which utilizes the familiar dynamic programming algorithm for calculating string-edit distance, and finding the best alignment between two arbitrary strings (Wagner and Fischer 1974). The algorithm finds subsets of the data that can be exactly-generated by sequential finite state automata of 3 and 4 states, as in (3), where the labels m_i should be understood as cover terms for morphemes in general. An automaton *exactly-generates* a set of strings S if it generates all strings in S and no other strings; a *sequential* FSA is one of the form sketched graphically in (1)-(3), where there is a unique successor to each state.

(3)

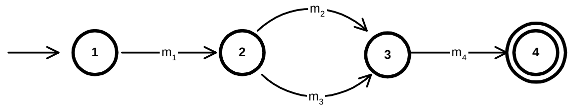


2.1 First stage: alignments.

If presented with the pair of strings *anapenda* and *anamupenda* from an unknown language, it is not difficult for a human being to come up with the hypothesis that *mu* is a morpheme inside a larger word that is composed of at least two morphemes, perhaps *ana-* and *-penda*. The SED heuristic makes this observation explicit by building small FSAs of the form in (4), where at most one of m_1 or m_4 may be null, and at most one of m_2 and m_3 may be null: we refer to these as *elementary alignments*. The strings m_2 and m_3 are called *counterparts*; the pairs of strings m_1 and m_4 are called the *context* (of the counterparts). (Indeed, we consider this kind of string comparison to be a plausible candidate for human language learning; see Dahan and Brent 1999).

[3] We use the terms *string prefix* and *string suffix* in the computer science sense: a string S is a string prefix of a string X iff there exists a string T such that X = S.T, where "." is the string concatenation operator; under such conditions, T is likewise a *string suffix* of X. Otherwise, we use the terms *prefix* and *suffix* in the linguistic sense, and a string prefix (e.g., *jump*) may be a linguistic stem, as in *jump-ing*.

(4)

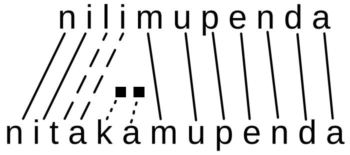


The first stage of the algorithm consists of looking at all pairs of words S, T in the corpus, and passing through the following steps:

We apply several initial heuristics to eliminate a large proportion of the pairs of strings before applying the familiar SED algorithm to them, in view of the relative slowness of the SED algorithm; see Goldsmith et al (2005) for further details.

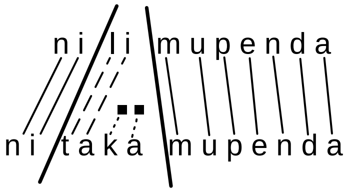
We compute the optimal alignment of S and T using the SED algorithm, where alignment between two identical letters (which we call *twins*) is assigned a cost of 0, alignment between two different letters (which we call *siblings*) is assigned a cost of 1.5, and a letter in one string not aligned with a segment on the other string (which we call an *orphan*) is assigned a cost of 1. An alignment as in (5) is thus assigned a cost of 5, based on a cost of 1.5 assigned to each broken line, and 1 to each dotted line that ends in a square box.

(5)



There is a natural map from every alignment to a unique sequence of pairs, where every pair is either of the form (S[i], T[j]) (representing either a twin or sibling case) or of the form (S[i], 0) or (0, T[j]) (representing the orphan case). We then divide the alignment up into perfect and imperfect spans: perfect spans are composed of maximal sequences of twin pairs, while imperfect spans are composed of maximal sequences of sibling or orphan pairs. This is illustrated in (6).

(6)



There is a natural equivalence between alignments and sequential FSAs as in (4), where perfect spans correspond to pairs of adjacent states with unique transitions and imperfect spans correspond to pairs of adjacent states with two transitions, and we will henceforth use the FSA notation to describe our algorithm.

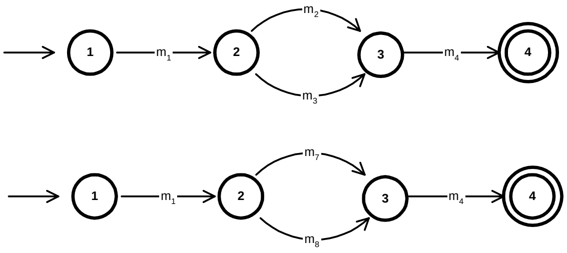
2.2 Collapsing alignments

As we noted above (4), for any elementary alignment, a *context* is defined: the pair of strings (one of them possibly null) which surround the pair of *counterparts*. Our first goal is to collapse alignments that share their context. We do this in the following way.

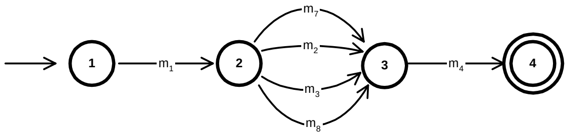
Let us define the set of strings associated with the paths leaving a state S as the *production* of state S. A four-state sequential FSA, as in (4), has three states with non-null productions; if this particular FSA corresponds to an *elementary alignment*, then two of the state-productions contain exactly one string—and these state-productions define the *context*— and one of the state-productions contains exactly two strings (one possibly the null string)—this defines the *counterparts*. If we have two such four-state FSAs whose context are identical, then we collapse the two FSAs into a single *conflated* FSA in which the context states and their productions are identical, and the states that produced the *counterparts* are collapsed by creating a state that produces the union of the productions of the original states. This is illustrated in (7): the two FSAs in (7a) share a context, generated by their states 1 and 3, and they are collapsed to form the FSA in (7b), in which the context states remain unchanged, and the counterpart states, labeled ‘2’, are collapsed to form a new state ‘2’ whose production is the union of the productions of the original states.

(7)

a.



b.



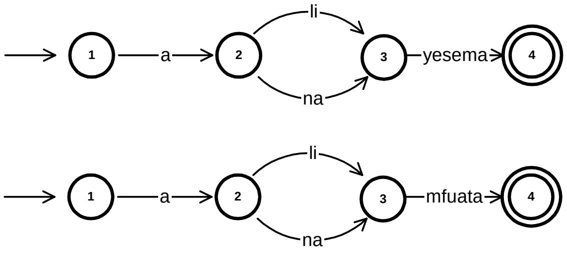
2.3 Collapsing the resulting sequential FSAs

We now generalize the procedure described in the preceding section to collapse any two sequential FSAs for which all but one of the corresponding states have exactly the same production. For example, the two sequential FSAs in (8a) are collapsed into (8b).

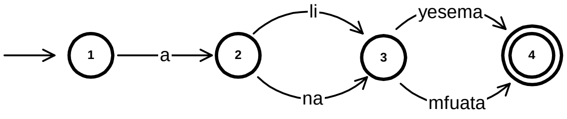
Three and four-state sequential FSAs as in (8b), where at least two of the state-transitions generate more than one morpheme, form the set of *templates* derived from our bootstrapping heuristic. Each such *template* can be usefully assigned a quantitative score based on the number of letters “saved” by the use of the template to generate the words, in the following sense. The template in (8b) summarizes four words: *aliyesema*, *alimfuata*, *anayesema*, and *anamfuata*. The total string length of these words is 36, while the total number of letters in the strings associated with the transitions in the FSA is 1+4+12 = 17; we say that the FSA *saves* 36-17 = 19 letters. In actual practice, the significant templates discovered save on the order of 200 to 5,000 letters, and ranking them by the number of letters saved is a good measure of how significant they are in the overall morphology of the language. We refer to this score as a template’s *robustness*; we employ this quantity again in section 3.1 below.

By this ranking, the top template found in our Swahili corpus of 50,000 running words was one that generated *a* and *wa* (class 1 and 2 subject markers) and followed by 246 correct verb continuations (all of them polymorphemic); the first 6 templates are summarized informally in Table 1. We note that the third and fourth template can also be collapsed to form a template of the form in (3), a point we return to below. Precision, recall, and F-score for these experiments are given in Table 2.

(8)
a.



b.



State 1	State 2	State 3
a, wa (sg., pl. human subject markers)	246 stems	
ku, hu (infinitive, habitual markers)	51 stems	
wa (pl. subject marker)	*ka, li* (tense markers)	25 stems
a (sg. subject marker)	*ka, li* (tense markers)	29 stems
a (sg. subject marker)	*ka, na* (tense markers	28 stems
37 strings	*w* (passive marker)	*a*

Table 1 Top templates in Swahili

	Precision	Recall	F-score
SED	0.77	0.57	0.65
SF	0.54	0.14	0.22
PF	0.68	0.20	0.31

Table 2 Results

3 Further developments

In this section, we describe three developments of the SED-based heuristic sketched in section 2. The first disambiguates which state it is that string material should be associated with in cases of ambiguity; the second *collapses* templates associated with similar morphological structure; the third uses the FSAs to *predict* words that do not actually occur in the corpus by hypothesizing stems on the basis of the established FSAs and as yet unanalyzed words in the corpus.

3.1 Disambiguating FSAs

In the case of a sequential FSA, when the final letter of the production of a (non-final) state S are identical, then that letter can be moved from being the string-suffix of all of the productions of state S to being the string-prefixes of all of the productions of the following state. More generally, when the *n* final letters of the productions of a state are identical, there is an n-way ambiguity in the analysis, and the same holds symmetrically for the ambiguity that arises when the *n* initial letters of the production of a (non-initial) state.

Thus two successive states, S and T, must (so to speak) fight over which will be responsible for generating the ambiguous string. We employ two steps to disambiguate these cases.

Step 1: The first step is applicable when the *number* of distinct strings associated with states S and T are quite different in size (typically corresponding to the case where one generates *grammatical* morphemes and the other generates *stems*); in this case, we assign the ambiguous material to the state that generates the smaller number of strings. There is a natural motivation for this choice from the perspective of our desire to minimize the size of the grammar, if we consider the size of the grammar to be based, in part, on the sum of the lengths of the morphemes produced by each state.

Step 2: It often happens that an ambiguity arises with regard to a string of one or more letters that could potentially be produced by either of a pair of successive states involving grammatical morphemes. To deal with this case, we make a decision that is also (like the preceding step) motivated by a desire to minimize the description length of the grammar. In this case, however, we think of the FSA as containing explicit *strings* (as we have assumed so far), but rather *pointers* to strings, and the “length” of a pointer to a string is inversely proportional to the logarithm of its frequency. Thus the *overall* use of a string in the grammar plays a crucial role in determining the length of a grammar, and we wish to maximize the appearance in our grammar of morphemes that are used frequently, and minimize the use of morphemes that are used rarely.

We implement this idea by collecting a table of all of the morphemes produced by our FSA, and assigning each a score which consists of the sum of the robustness scores of each template they occur in (see discussion just above (8)). Thus morphemes occurring in several high robustness templates will have high scores; morphemes appearing in a small number of lowly ranked templates will have low scores.

To disambiguate strings which could be produced by either of two successive states, we consider all possible parsings of the string between the states, and score each parsing as the sum of the scores of the component morphemes; we chose the parsing for which the total score is a maximum.

For example, Swahili has two common tense markers, *ka* and *ki*, and this step corrected a template from {*ak*}+{*a*,*i*}+{stems} to {*a*}+{*ka*,*ki*}+{stems}, and others of similar form. It also did some useful splitting of joined morphemes, as when it modified a template {*wali*} + {NULL, *po*} + {stems} to {*wa*} + {*li*, *lipo*} + {stems}. In this case, *wali* should indeed be split into *wa* + *li* (subject and tense markers, resp.), and while the change creates an error (in the sense that *lipo* is, in fact, two morphemes; *po* is a subordinate clause marker), the resulting error occurs considerably less often in the data, and the correct template will better be able to be integrated with out templates.

3.2 Template collapsing

From a linguistic point of view, the SED-based heuristic creates too many FSAs because it stays too close to the data provided by the corpus. The only way to get a more correct grammar is by collapsing the FSAs, which will have as a consequence the generation of new words not found in the corpus. We apply the following relatively conservative strategy for collapsing two templates.

We compare templates of the same number of states, and distinguish between states that produce grammatical morphemes (five or fewer in number) and those that produce stems (that is, lexical morphemes, identified as being six or more in number). We collapse two templates if the productions of the corresponding states satisfy the following conditions: if the states generate stems, then the intersection of the productions must be at least two stems, while if the states are grammatical morphemes, then the productions of one pair of corresponding states must be *identical*, while for the other pair, the symmetric difference of the productions must be no greater than two in number (that is, the number of morphemes produced by the state of one template but not the other must not exceed 2).

3.3 Reparsing words in the corpus and predicting new words

When we create robust FSAs—that is, FSAs that generate a large number of words—we are in a position to go back to the corpus and reanalyze a large number of words that could not be previously analyzed. That is, a 4-state FSA in which each state produced two strings generates 8 words, and *all* 8 words must appear in the corpus for the method described so far in order for this particular FSA to generate *any* of them. But that condition is unlikely to be satisfied for any but the most common of morphemes, so we need to go back to the corpus and *infer* the existence of new stems (as defined operationally in the preceding paragraph) based on their occurrence in several, but not *all* possible, words. If there exist 3 distinct words in the corpus which would all be generated by a template if a given stem were added to the template, we add that stem to the template.

4 Experiments and Results

In this section, we present three sets of evaluations of the refinements of the SED heuristics described in the preceding section. We used a corpus of 7,180 distinct words occurring in 50,000 running words from a Swahili translation of the Bible obtained on the internet.

4.1 Disambiguating FSAs

In order to evaluate the effects of the disambiguating of FSAs described in section 3.1, we compare precision and recall of the identification of morpheme boundaries using the SED method *with* and *without* the disambiguation procedure described above. In Figures 1 and 2, we graph precision and recall for the top 10% of the templates, displayed as the leftmost point, for the top 20% of the templates, displayed as the second point from the left; and so on, because the higher ranked FSAs are more intrinsically more reliable than the lower ranked ones. We see that disambiguation repairs almost 50% of the previous errors, and increases recalls by about 10%. With these increases in precision and recall, it is clear that the disambiguating step provides a considerably more accurate morpheme boundary discovery procedure.

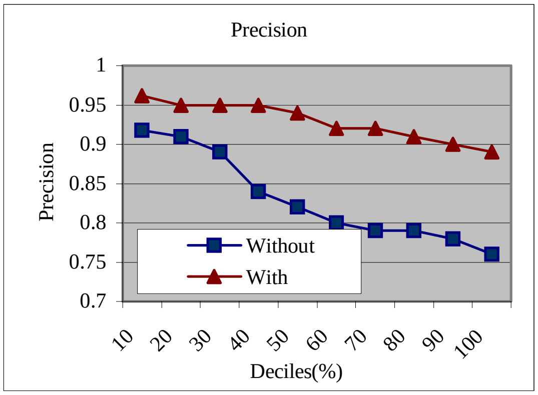


Figure 1 Comparison of precision

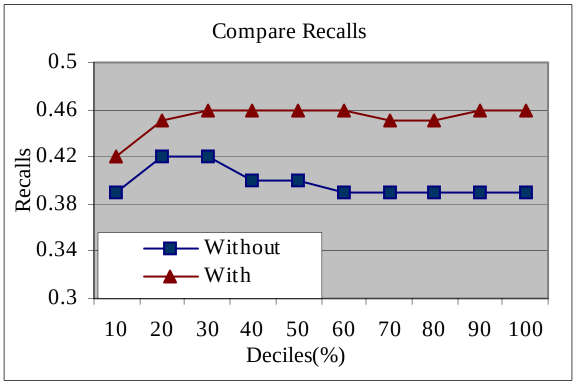


Figure 2 Comparison of recall

4.2 Template collapsing

The second refinement discussed above consists of finding pairs of similar templates, collapsing them as appropriate, and thus creating patterns that generate new words that did not participate in the formation of the original templates. These new words may or may not themselves appear in the corpus. We are, however, able to judge their morphological well-formedness by inspection. We list in Table 3 the entire list of eight templates that are collapsed in this step.

All of the templates which are collapsed in this step are in fact of the same morphological structure (with one very minor exception[4]): they are of the form *subject marker + tense marker + stem,* and the collapsing induced in this procedure correctly creates larger templates of precisely the same structure, generating new words not seen in the corpus that are in fact correct from our (non-native speaker) inspection. We submitted the new words to Yahoo to test the words "existence" by their existence on the internet, and actually found an average of 87% of the predicted words in a template; see the last column in Table 3 for details.

4.3 Reparsing

After previous refinements, we obtain a number of robust FSAs, for example, those collapsed templates in Table 3. With them, we then search the corpus for those words that can only be partly fitted into these FSAs and generate associated stems. Table 4 shows the reparsed words that had not been parsed by earlier templates and also newly added stems for some robust FSAs (the four collapsed templates in Table 3). Stems such as *anza* 'begin' and *fanya* 'do' are thus added to the first template, and all words derived by prepending a tense marker and a subject marker are indeed accurate words. As the words in Table 4 suggest, the reparsing process adds new, common stems to the stem-column of the templates, thus making it easier for the collapsing function to find similarities across related templates.

In future work, we will take use the larger templates, populated with more stems, and input them to the collapsing function described in 3.2.

5 Conclusions

On the basis of the experiments with Swahili described in this paper, the SED heuristic appears to be a useful tool for the discovery of morphemes in languages with rich morphologies, and for the discovery of the FSAs that constitute the morphologies of those languages.

Ultimately, the value of the heuristic is best tested against a range of languages with complex concatenative morphologies. While a thorough discussion would take us well beyond the limits of this paper, we have applied the SED heuristic to English, Hungarian, and Finnish as well as Swahili. For English, unsurprisingly, the method works as well as the SF and PF methods, though a bit more slowly, while for Hungarian and Finnish, the results appear promising, and a comparison with Creutz and Lagus (2004) for Finnish, for example, would be appealing.

[4] The exception involves the distinct morpheme *po,* a subordinate clause marker which must ultimately be analyzed as appearing in a distinct template column to the right of the tense markers.

	One Template	The other template	Collapsed Template	% found on Yahoo search
1	{a}-{ka,na}-{stems}	{a}-{ka,ki}-{stems}	{a}-{ka,ki,na}-{stems}	86 (37/43)
2	{wa}-{ka,na}-{stems}	{wa}-{ka,ki}-{stems}	{wa}-{ka,ki,na}-{stems}	95 (21/22)
3	{a}-{ka,ki,na}-{stems}	{wa}-{ka,ki,na}-{stems}	{a,wa}-{ka,ki,na}-{stems}	84 (154/183)
4	{a}-{liye,me}-{stems}	{a}-{liye,li}-{stems}	{a}-{liye,li,me}-{stems}	100 (21/21)
5	{a}-{ki,li}-{stems}	{wa}-{ki,li}-{stems}	{a,wa}-{ki,li}-{stems}	90 (36/40)
6	{a}-{lipo,li}-{stems}	{wa}-{lipo,li}-{stems}	{a,wa}-{lipo,li}-{stems}	90 (27/30)
7	{a,wa}-{ki,li}-{stems}	{a,wa}-{lipo,li}-{stems}	{a,wa}-{ki,lipo,li}-{stems}	74 (52/70)
8	{a}-{na,naye}-{stems}	{a}-{na,ta}-{stems}	{a}-{na,ta,naye}-{stems}	80 (12/15)

Table 3 Collapsed Templates and Created Words Sample.

	Template	Reparsed Words Not Parsed Before	Added Stems
1	{a, wa}-{ka,ki,na}-{stems}	akawakweza, akiwa, anacho, akibatiza, …	toka, anza, waita, fanya, enda, …
2	{a}-{li,liye,me }-{stems}	ameinuka, ameugua, alivyo, aliyoniagiza, …	zaliwa, kuwa, fanya, sema
3	{a, wa}-{ki,li,lipo}-{stems}	alimtoboa, alimtaka, waliamini, …	pata, kuwa, kaa, fanya, chukua, fika, …
4	{a} – {na,naye,ta}-{stems}	analazwa, atanitukuza, anaye, anakuita, …	ingia, sema

Table 4 Reparsed words and "discovered" stems

References

Creutz, Mathias, and Krista Lagus. (2004). Induction of a simple morphology for highly inflecting languages. *Proceedings of the Workshop of SIGPHON* (Barcelona).

Dahan, Delphine, and Michael Brent. (1999). On the discovery of novel world-like units from utterances. *Journal of Experimental Psychology: General* 128: 165-185.

Goldsmith, John. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2): 153-198.

Goldsmith, John, Yu Hu, Irina Matveeva, and Colin Sprague. 2005. A heuristic for morpheme discovery based on string edit distance. Technical Report TR-2005-4. Department of Computer Science. University of Chicago.

Hafer, M. A., Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10: 371-385.

Harris, Zellig. (1955). From Phoneme to Morpheme. *Language* 31: 190-222.

Harris, Zellig. (1967). Morpheme Boundaries within Words: Report on a Computer Test. *Transformations and Discourse Analysis Papers* 73.

Oliver, Antoni, Irene Bastellón, and Lluís Màrquez. (2003). Uso de Internet para aumentar la cobertura de un sistema de adquisición léxica del ruso. SEPLN 2003.

Wagner, R. A., Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery* 21(1): 168-73.

van Zaanen, Menno. 2000. ABL: Alignment-Based Learning. *Proceedings of the 17th Conference on Computational Linguistics*, vol. 2. p. 961-67.

A Connectionist Model of Language-Scene Interaction

Marshall R. Mayberry, III   **Matthew W. Crocker**   **Pia Knoeferle**
Department of Computational Linguistics
Saarland University
Saarbrücken 66041, Germany
martym,crocker,knoferle@coli.uni-sb.de

Abstract

Recent "visual worlds" studies, wherein researchers study language in context by monitoring eye-movements in a visual scene during sentence processing, have revealed much about the interaction of diverse information sources and the time course of their influence on comprehension. In this study, five experiments that trade off scene context with a variety of linguistic factors are modelled with a Simple Recurrent Network modified to integrate a scene representation with the standard incremental input of a sentence. The results show that the model captures the qualitative behavior observed during the experiments, while retaining the ability to develop the correct interpretation in the absence of visual input.

1 Introduction

People learn language within the context of the surrounding world, and use it to refer to objects in that world, as well as relationships among those objects (e.g., Gleitman, 1990). Recent research in the *visual worlds* paradigm, wherein participants' gazes in a scene while listening to an utterance are monitored, has yielded a number of insights into the time course of sentence comprehension. The careful manipulation of information sources in this experimental setting has begun to reveal important characteristics of comprehension such as incrementality and anticipation. For example, people's attention to objects in a scene closely tracks their mention in a spoken sentence (Tanenhaus et al., 1995), and world and linguistic knowledge seem to be factors that facilitate object identification (Altmann and Kamide, 1999; Kamide et al., 2003). More recently, Knoeferle et al. (2005) have shown that when scenes include depicted events, such visual information helps to establish important relations between the entities, such as role relations.

Models of sentence comprehension to date, however, continue to focus on modelling reading behavior. No model, to our knowledge, attempts to account for the use of immediate (non-linguistic) context. In this paper we present results from two simulations using a Simple Recurrent Network (SRN; Elman, 1990) modified to integrate input from a scene with the characteristic incremental processing of such networks in order to model people's ability to adaptively use the contextual information in visual scenes to more rapidly interpret and disambiguate a sentence. In the modelling of five visual worlds experiments reported here, accurate sentence interpretation hinges on proper case-role assignment to sentence referents. In particular, modelling is focussed on the following aspects of sentence processing:

- anticipation of upcoming arguments and their roles in a sentence
- adaptive use of the visual scene as context for a spoken utterance
- influence of depicted events on developing interpretation
- multiple/conflicting information sources and their relative importance

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 36–44,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

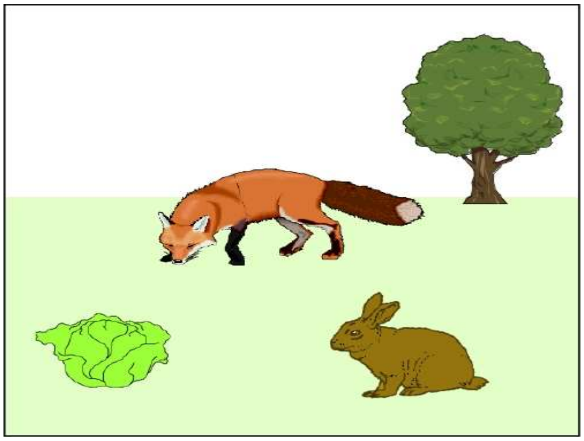


Figure 1: **Selectional Restrictions.** Gaze fixations depend on whether the hare is the subject or object of the sentence, as well as the thematic role structure of the verb. These gaze fixations reveal that people use linguistic and world knowledge to anticipate upcoming arguments.

2 Simulation 1

In the first simulation, we simultaneously model four experiments that featured revealing contrasts between world knowledge and context. These four experiments show that the human sentence processor is very adept at utilizing all available sources of information to rapidly interpret language. In particular, information from visual context can readily be integrated with linguistic and world knowledge to disambiguate argument roles where the information from the auditory stream is insufficient in itself.

All experiments were conducted in German, a language that allows both subject-verb-object (SVO) and object-verb-subject (OVS) sentence types, so that word order alone cannot be relied upon to determine role assignments. Rather, case marking in German is used to indicate grammatical function such as subject or object, except in the case of feminine and neuter nouns where the article does not carry any distinguishing marking for the nominative and accusative cases.

2.1 Anticipation depending on stereotypicality

The first two experiments modelled involved unambiguous sentences in which case-marking and verb selectional restrictions in the linguistic input (i.e., linguistic and world knowledge or stereotypicality), together with characters depicted in a visual scene, allowed rapid assignment of the roles played by those characters in the sentence.

Experiment 1: Morphosyntactic and lexical verb information. In order to examine the influence of linguistic knowledge of case-marking, Kamide et al. (2003) presented experiment participants with a scene showing, for example, a hare, a cabbage, a fox, and a distractor (see Figure 1), together with either a spoken German SVO sentence (1) or with an OVS sentence (2):

(1) *Der Hase frisst gleich den Kohl.*
The hare$_{nom}$ eats shortly the cabbage$_{acc}$.

(2) *Den Hasen frisst gleich der Fuchs.*
The hare$_{acc}$ eats shortly the fox$_{nom}$.

The subject and object case-marking on the article of the first noun phrase (NP) together with verb meaning and world knowledge allowed anticipation of the correct post-verbal referent. Participants made anticipatory eye-movements to the cabbage after hearing "The hare$_{nom}$ eats ..." and to the fox upon encountering "The hare$_{acc}$ eats ...". Thus, people are able to predict upcoming referents when the utterance is unambiguous and linguistic/world knowledge restricts the domain of potential referents in a scene.

Experiment 2: Verb type information. To further investigate the role of verb information, the authors used the same visual scenes in a follow-up study, but replaced the agent/patient verbs like *frisst* ("eats") with experiencer/theme verbs like *interessiert* ("interests"). The agent (experiencer) and patient (theme) roles from Experiment 1 were interchanged. Given the same scene in Figure 1 but the subject-first sentence (3) or object-first sentence (4), participants showed gaze fixations complementary to those in the first experiment, confirming that both syntactic case information and semantic verb information are used to predict subsequent referents.

(3) *Der Hase interessiert ganz besonders den Fuchs.*
The hare$_{nom}$ interests especially the fox$_{acc}$.

(4) *Den Hasen interessiert ganz besonders der Kohl.*
The hare$_{acc}$ interests especially the cabbage$_{nom}$.

2.2 Anticipation depending on depicted events

The second set of experiments investigated temporarily ambiguous German sentences. Findings showed that depicted events–just like world and linguistic knowledge in unambiguous sentences–can establish a scene character's role as agent or patient in the face of linguistic structural ambiguity.

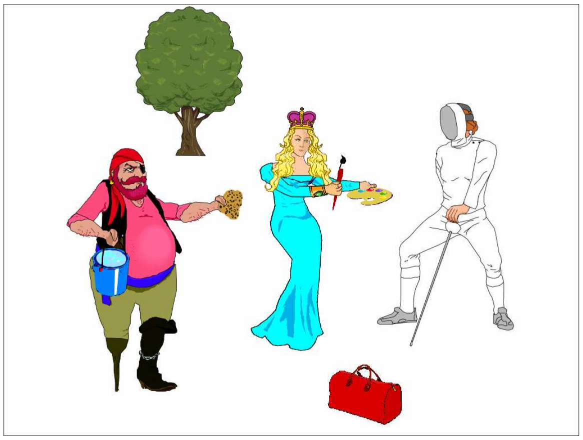


Figure 2: **Depicted Events.** The depiction of actions allows role information to be extracted from the scene. People can use this information to anticipate upcoming arguments even in the face of ambiguous linguistic input.

Experiment 3: Verb-mediated depicted role relations. Knoeferle et al. (2005) investigated comprehension of spoken sentences with local structural and thematic role ambiguity. An example of the German SVO/OVS ambiguity is the SVO sentence (5) versus the OVS sentence (6):

(5) *Die Princessin malt offensichtlich den Fechter*.
The princess$_{nom}$ paints obviously the fencer$_{acc}$.

(6) *Die Princessin wäscht offensichtlich der Pirat.*
The princess$_{acc}$ washes obviously the pirate$_{nom}$.

Together with the auditorily presented sentence a scene was shown in which a princess both paints a fencer and is washed by a pirate (see Figure 2). *Linguistic* disambiguation occurred on the second NP; in the absence of stereotypical verb-argument relationships, disambiguation prior to the second NP was only possible through use of the depicted events and their associated depicted role relations. When the verb identified an action, the depicted role relations disambiguated towards either an SVO agent-patient (5) or OVS patient-agent role (6) relation, as indicated by anticipatory eye-movements to the patient (pirate) or agent (fencer), respectively, for (5) and (6). This gaze-pattern showed the rapid influence of verb-mediated depicted events on the assignment of a thematic role to a temporarily ambiguous sentence-initial noun phrase.

Experiment 4: Weak temporal adverb constraint. Knoeferle et al. also investigated German verb-final active/passive constructions. In both the active future-tense (7) and the passive sentence (8), the initial subject noun phrase is role-ambiguous, and the auxiliary *wird* can have a passive or future interpretation.

(7) *Die Princessin wird sogleich den Pirat washen*.
The princess$_{nom}$ will right away wash the pirate$_{acc}$.

(8) *Die Princessin wird soeben von dem Fechter gemalt.*
The princess$_{acc}$ is just now painted by the fencer$_{nom}$.

To evoke early linguistic disambiguation, temporal adverbs biased the auxiliary *wird* toward either the future ("will") or passive ("is -ed") reading. Since the verb was sentence-final, the interplay of scene and linguistic cues (e.g., temporal adverbs) were rather more subtle. When the listener heard a future-biased adverb such as *sogleich*, after the auxiliary *wird*, he interpreted the initial NP as an agent of a future construction, as evidenced by anticipatory eye-movements to the patient in the scene. Conversely, listeners interpreted the passive-biased construction with these roles exchanged.

2.3 Architecture

The Simple Recurrent Network is a type of neural network typically used to process temporal sequences of patterns such as words in a sentence. A common approach is for the modeller to train the network on prespecified targets, such as verbs and their arguments, that represent what the network is expected to produce upon completing a sentence. Processing is incremental, with each new input word interpreted in the context of the sentence processed so far, represented by a copy of the previous hidden layer serving as additional input to the current hidden layer. Because these types of associationist models automatically develop correlations among the sentence constituents they are trained on, they will generally develop expectations about the output even before processing is completed because sufficient information occurs early in the sentence to warrant such predictions. Moreover, during the course of processing a sentence these expectations can be overridden with subsequent input, often abruptly revising an interpretation in a manner reminiscent of how humans seem to process language. Indeed, it is these characteristics of incremental processing, the automatic development of expectations, seamless integration of multiple sources of information, and nonmonotonic revision that have endeared neural network models to cognitive researchers.

In this study, the four experiments described

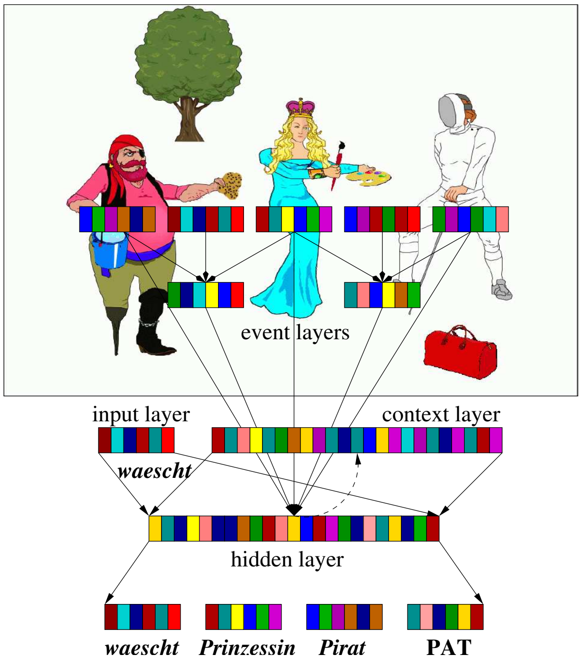


Figure 3: **Scene Integration.** A simple conceptual representation of the information in a scene, along with compressed event information from depicted actions when present, is fed into a standard SRN to model adaptive processing. The links connecting the depicted characters to the hidden layer are shared, as are the links connecting the event layers to the hidden layer.

above have been modelled simultaneously using a single network. The goal of modelling all experimental results by a single architecture required enhancements to the SRN, the development and presentation of the training data, as well as the training regime itself. These will be described in turn below.

In two of the experiments, only three characters are depicted, representation of which can be propagated directly to the network's hidden layer. In the other two experiments, the scene featured three characters involved in two events (e.g., **pirate-washes-princess** and **princess-paints-fencer**, as shown in Figure 3). The middle character was involved in both events, either as an agent or a patient (e.g., **princess**). Only one of the events, however, corresponded to the spoken linguistic input.

The representation of this scene information and its integration into the model's processing was the main modification to the SRN. Connections between representations for the depicted characters and the hidden layer were provided. Encoding of the depicted events, when present, required additional links from the characters and depicted actions to **event** layers, and links from these event layers to the SRN's hidden layer. The network developed representations for the events in the event layers by compressing the scene representations of the involved characters and depicted actions through weights corresponding to the action, its agent and its patient for each event. This event representation was kept simple and only provided conceptual input to the hidden layer: who did what to whom was encoded for both events, when depicted, but grammatical information only came from the linguistic input. As the focus of this study was on whether sentence processing could adapt to information from the scene when present or from stored knowledge, lower-level perceptual processes such as attention were not modelled.

Neural networks will usually encode any correlations in the data that help to minimize error. In order to prevent the network from encoding regularities in its weights regarding the position of the characters and events given in the scene (such as, for example, that the central character in the scene corresponds to the first NP in the presented sentence) which are not relevant to the role-assignment task, one set of weights was used for all characters, and another set of weights used for both events. This weight-sharing ensured that the network had to access the information encoded in the event layers, or determine the relevant characters itself, thus improving generalization. The representations for the characters and actions were the same for both input (scene and sentence) and output.

The input assemblies were the scene representations and the current word from the input sentence. The output assemblies were the verb, the first and second nouns, and an assembly that indicated whether the first noun was the agent or patient of the sentence (token **PAT** in Figure 3). Typically, agent and patient assemblies would be fixed in a case-role representation without such a discriminator, and the model required to learn to instantiate them correctly (Miikkulainen, 1997). However, we found that the model performed much better when the task was recast as having to learn to isolate the nouns in the order in which they are introduced, and separately mark how those nouns relate to the verb. The input and output assemblies had 100 units each, the event layers contained 200 units each, and the hidden and context layers consisted of 400 units.

2.4 Input Data, Training, and Experiments

We trained the network to correctly handle sentences involving non-stereotypical events as well as stereotypical ones, both when visual context was present and when it was absent. As over half a billion sentence/scene combinations were possible for all of the experiments, we adopted a grammar-based approach to exhaustively generate sentences and scenes based on the experimental materials while holding out the actual materials to be used for testing. In order to accurately model the first two experiments involving selectional restrictions on verbs, two additional words were added to the lexicon for each character selected by a verb. For example, in the sentence *Der Hase frisst gleich den Kohl*, the nouns *Hase1*, *Hase2*, *Kohl1*, and *Kohl2* were used to develop training sentences. These were meant to represent, for example, words such as “rabbit” and “jackrabbit” or “carrot” and “lettuce” in the lexicon that have the same distributional properties as the original words “hare” and “cabbage”. With these extra tokens the network could learn that *Hase*, *frisst*, and *Kohl* were correlated without ever encountering all three words in the same training sentence. The experiments involving non-stereotypicality did not pose this constraint, so training sentences were simply generated to avoid presenting experimental items.

Some standard simplifications to the words have been made to facilitate modelling. For example, multi-word adverbs such as *fast immer* were treated as one word through hyphenation so that sentence length within a given experimental set up is maintained. Nominal case markings such as *-n* in *Hasen* were removed to avoid sparse data as these markings are idiosyncratic, while the case markings on the determiners are more informative overall. More importantly, morphemes such as the infinitive marker *-en* and past participle *ge-* were removed, because, for example, the verb forms *malt*, *malen*, and *gemalt*, would all be treated as unrelated tokens, again contributing unnecessarily to the problem with sparse data. The result is that one verb form is used, and to perform accurately, the network must rely on its position in the sentence (either second or sentence-final), as well as whether the word *von* occurs to indicate a participial reading rather than infinitival. All 326 words in the lexicon for the first four experiments were given random representations over the vertices of a 100-dimensional hypercube, which resulted in marked improvement over sampling from within the hypercube (Noelle et al., 1997).

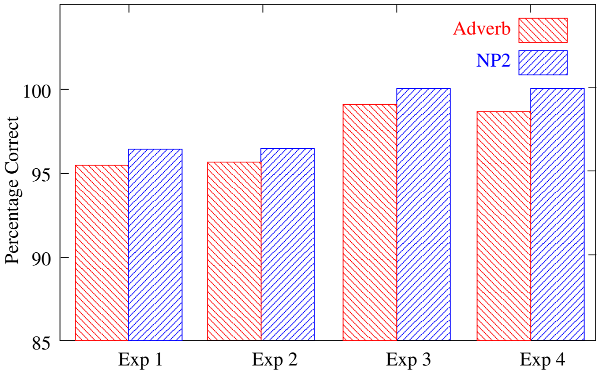


Figure 4: **Results.** In each of the four experiments modelled, anticipation of the upcoming argument at the adverb is nearly as accurate as at sentence end. However, the network has some difficulty with distinguishing stereotypical arguments.

We trained the network by repeatedly presenting the model with 1000 randomly generated sentences from each experiment (constituting one epoch) and testing every 100 epochs against the held-out test materials for each of the four experiments. Scenes were provided half of the time to provide an unbiased approximation to linguistic experience. The network was initialized with weights between -0.01 and 0.01. The learning rate was initially set to 0.05 and gradually reduced to 0.002 over the course of 15000 epochs. Ten splits were run on 1.6Ghz PCs and took a little over two weeks to complete.

2.5 Results

Figure 4 reports the percentage of targets at the network’s output layer that the model correctly matches, both as measured at the adverb and at the end of the sentence. The model clearly demonstrates the qualitative behavior observed in all four experiments in that it is able to access the information from the encoded scene or stereotypicality and combine it with the incrementally presented sentence to anticipate forthcoming arguments.

For the two experiments (1 and 2) using stereotypical information, the network achieved just over 96% at sentence end, and anticipation accuracy was just over 95% at the adverb. Analysis shows that the network makes errors in token identification, confusing words that are within the selectionally restricted set, such as, for example, *Kohl* and *Kohl2*. Thus, the model has not quite mastered the stereotypical knowledge, particularly as it relates to the presence of the scene.

For the other two experiments using non-stereotypical characters and depicted events (experiments 3 and 4), accuracy was 100% at the end of the sentence. More importantly, the model achieved over 98% early disambiguation on experiment 3, where the sentences were simple, active SVO and OVS. Early disambiguation on experiment 4 was somewhat harder because the adverb is the disambiguating point in the sentence as opposed to the verb in the other three experiments. As nonlinear dynamical systems, neural networks sometimes require an extra step to settle after a decision point is reached due to the attractor dynamics of the weights.

On closer inspection of the model's behavior during processing, it is apparent that the event layers provide enough additional information beyond that encoded in the weights between the characters and the hidden layer that the model is able to make finer discriminations in experiments 3 and 4, enhancing its performance.

3 Simulation 2

The previous set of experiments examined how people are able to use either stereotypical knowledge or depicted information to anticipate forthcoming arguments in a sentence. But how does the human sentence processor handle these information sources when both are present? Which takes precedence when they conflict? The experiment modelled in this section was designed to provide some insight into these questions.

Scene vs Stored Knowledge. Based on the findings from the four experiments in Simulation 1, Knoeferle and Crocker (2004b) examined two issues. First, it verified that stored knowledge about non-depicted events and information from depicted, but non-stereotypical, events each enable rapid thematic interpretation. An example scene showed a wizard, a pilot, and a detective serving food (Figure 5). When people heard condition 1 (example sentence 9), the case-marking on the first NP identified the pilot as a patient. Stereotypical knowledge identified the wizard as the only relevant agent, as indicated by a higher proportion of anticipatory eye-movements to the stereotypical agent (wizard) than to the detective. In contrast, when people heard the verb in condition 2 (sentence 10), it uniquely identified the detective as the only food-serving agent, revealed by more inspections to the agent of the depicted event (detective) than to the wizard.

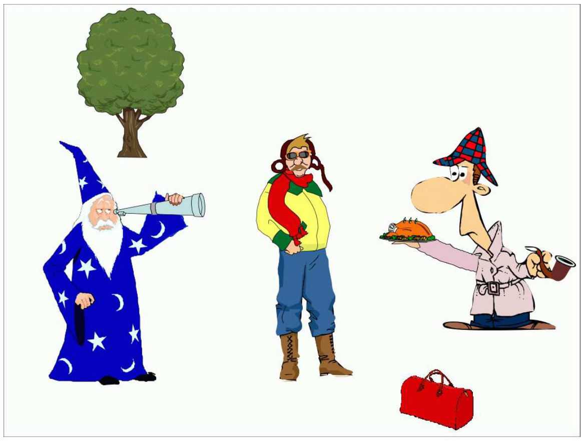


Figure 5: **Scene vs Stored Knowledge.** Experimental results show that people rely on depicted information over stereotypical knowledge when both are present during sentence processing.

(9) *Den Piloten verzaubert gleich der Zauberer*.
The pilot$_{acc}$ jinxes shortly the wizard$_{nom}$.

(10) *Den Piloten verköstigt gleich der Detektiv*.
The pilot$_{acc}$ serves-food-to shortly the detective$_{nom}$.

Second, the study determined the *relative importance* of depicted events and verb-based thematic role knowledge when the information sources were in competition. In both conditions 3 & 4 (sentences 11 & 12), participants heard an utterance in which the verb identified both a stereotypical (detective) and a depicted agent (wizard). When faced with this conflict, people preferred to rely on the immediate event depictions over stereotypical knowledge, and looked more often at the wizard, the agent in the depicted event, than at the other, stereotypical agent of the spying-action (the detective).

(11) *Den Piloten bespitzelt gleich der Detektiv*.
The pilot$_{acc}$ spies-on shortly the detective$_{nom}$.

(12) *Den Piloten bespitzelt gleich der Zauberer*.
The pilot$_{acc}$ spies-on shortly the wizard$_{nom}$.

3.1 Architecture, Data, Training, and Results

In simulation 1, we modelled experiments that depended on stereotypicality or depicted events, but not both. The experiment modelled in simulation 2, however, was specifically designed to investigate how these two information sources interacted. Accordingly, the network needed to learn to use either information from the scene or stereotypicality when available, and, moreover, favor the scene when the two sources conflicted, as observed in the empirical results. Recall that the network is trained only on the final interpretation of a sentence. Thus, capturing the observed behavior required manipulation of the frequencies of the four conditions described above during training. In order to train the network to develop stereotypical agents for verbs, the frequency that a verb occurs with its stereotypical agent, such as *Detektiv* and *bespitzelt* from example (11) above, had to be greater than for a non-stereotypical agent. However, the frequency should not be so great that it overrode the influence from the scene.

The solution we adopted is motivated by a theory of language acquisition that takes into account the importance of early linguistic experience in a visual environment (see the General Discussion). We found a small range of ratios of stereotypicality to non-stereotypicality that permitted the network to develop an early reliance on information from the scene while it gradually learned the stereotypical associations. When the ratio was lower than 6:1, the network developed too strong a reliance on stereotypicality, overriding information from the scene. When the ratio was greater than 15:1, the scene always took precedence when it was present, but stereotypical knowledge was used when the scene was not present. Within this range, however, the network quickly learns to extract information from the scene because the scene representation remains static while a sentence is processed incrementally. It is the stereotypical associations, predictably, that take longer for the network to learn in rough proportion to their ratio over non-stereotypical agents.

Figure 6 shows the effect this training regime had over 6000 epochs on the ability of the network to accurately anticipate the missing argument in each of the four conditions described above when the ratio of non-stereotypical to stereotypical sentences was 8:1. The network quickly learns to use the scene for conditions 2-4 (examples 10-12), where the action in the linguistic input stream is also depicted, allowing the network to determine the relevant event and deduce the missing argument. (Because conditions 3 and 4 are the same up to the second NP, their curves are, in fact, identical.) But condition 1 (sentence 9) requires only stereotypical knowledge. The accuracy of condition 1 remains close to 75% (correctly producing the verb, first NP, and role discriminator, but not the second NP) until around epoch 1200 or so and then gradually improves as the network learns the appropriate stereotypical associations. The condition 1 curve asymptotically approaches 100% over the course of 10,000 epochs.

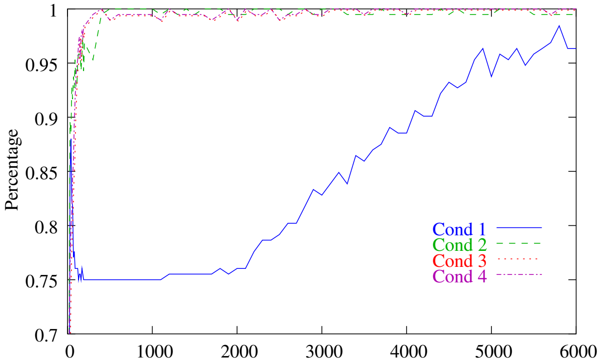


Figure 6: **Acquisition of Stereotypicality.** Stereotypical knowledge (condition 1) is acquired much more gradually than information from the scene (conditions 2-4).

Results from several runs with different training parameters (such as learning rate and stereotypicality ratio) show that the network does indeed model the observed experimental behavior. The best results so far exceed 99% accuracy in correctly anticipating the proper roles and 100% accuracy at sentence end.

As in simulation 1, the training corpus was generated by exhaustively combining participants and actions for all experimental conditions while holding out all test sentences. However, we found that we were able to use a larger learning rate, 0.1, than the 0.05 used in the first simulation. The 130 words in the lexicon were given random binary representations from the vertices of a 100-dimensional hypercube as described before.

Analysis of the network after successful training suggests why the training regime of holding the ratio of stereotypical to non-stereotypical sentences constant works. Early in training, before stereotypicality has been encoded in the network's weights, patterns are developed in the hidden layer as each word is processed that enable the network to accurately decode the words in the output layer. Once the verb is read in, its hidden layer pattern is available to produce the correct output representations for both the verb itself and its stereotypical agent. Not surprisingly, the network thus learns to associate the hidden layer pattern for the verb with its stereotypical agent pattern in the second NP output slot. The only constraint for the network is to ensure that the scene can still override this stereotypicality when the depicted event so dictates.

4 General Discussion and Future Work

Experiments in the visual worlds paradigm have clearly reinforced the view of language comprehension as an active, incremental, highly integrative process in which anticipation of upcoming arguments plays a crucial role. Visual context not only facilitates identification of likely referents in a sentence, but helps establish relationships between referents and the roles they may fill. Research thus far has shown that the human sentence processor seems to have easy access to whatever information is available, whether it be syntactic, lexical, semantic, or visual, and that it can combine these sources to achieve as complete an interpretation as is possible at any given point in comprehending a sentence.

The modelling results reported in this paper are an important step toward the goal of understanding how the human sentence processor is able to accomplish these feats. The SRN provides a natural framework for this research because its operation is premised on incremental and integrative processing. Trained simply to produce a representation of the complete interpretation of a sentence as each new word is processed (on the view that people learn to process language by reviewing what they hear), the model automatically develops anticipations for upcoming arguments that allow it to demonstrate the early disambiguation behavior observed in the visual worlds experiments modelled here.

The simple accuracy results belie the complexity of the task in both simulations. In Simulation 1, the network has to demonstrate early disambiguation when the scene is present, showing that it can indeed access the proper role and filler from the compressed representation of the event associated with the first NP and verb processed in the linguistic stream. This task is rendered more difficult because the proper event must be extracted from the superimposition of the two events in the scene, which is what is propagated into the model's hidden layer. In addition, it must also still be able to process all sentences correctly when the scene is not present.

Simulation 2 is more difficult still. The experiment shows that information from the scene takes precedence when there is a conflict with stereotypical knowledge; otherwise, each source of knowledge is used when it is available. In the training regime used in this simulation, the dominance of the scene is established early because it is much more frequent than the more particular stereotypical knowledge. As training progresses, stereotypical knowledge is gradually learned because it is sufficiently frequent for the network to capture the relevant associations. As the network weights gradually saturate, it becomes more difficult to retune them. But encoding stereotypical knowledge requires far fewer weight adjustments, so the network is able to learn that task later during training.

Knoeferle and Crocker (2004a,b) suggest that the preferred reliance of the comprehension system on the visual context over stored knowledge might best be explained by appealing to a boot-strapping account of language acquisition such as that of Gleitman (1990). The development of a child's world knowledge occurs in a visual environment, which accordingly plays a prominent role during language acquisition. The fact that the child can draw on two informational sources (utterance and scene) enables it to infer information that it has not yet acquired from what it already knows. This contextual development may have shaped both our cognitive architecture (i.e., providing for rapid, seamless integration of scene and linguistic information), and comprehension mechanisms (e.g., people rapidly avail themselves of information from the immediate scene when the utterance identifies it).

Connectionist models such as the SRN have been used to model aspects of cognitive development, including the timing of emergent behaviors (Elman et al., 1996), making them highly suitable for simulating developmental stages in child language acquisition (e.g., first learning names of objects in the immediate scene, and later proceeding to the acquisition of stereotypical knowledge). If there are developmental reasons for the preferred reliance of listeners on the immediate scene during language comprehension, then the finding that modelling that development provides the most efficient (if not only) way to naturally reproduce the observed experimental behavior promises to offer deeper insight into how such knowledge is instilled in the brain.

Future research will focus on combining all of the experiments in one model, and expand the range of sentence types and fillers to which the network is exposed. The architecture itself is being redesigned to scale up to much more complex linguistic constructions and have greater coverage while retaining the cognitively plausible behavior described in this study (Mayberry and Crocker, 2004).

5 Conclusion

We have presented a neural network architecture that successfully models the results of five recent experiments designed to study the interaction of visual context with sentence processing. The model shows that it can adaptively use information from the visual scene such as depicted events, when present, to anticipate roles and fillers as observed in each of the experiments, as well as demonstrate traditional incremental processing when context is absent. Furthermore, more recent results show that training the network in a visual environment, with stereotypical knowledge gradually learned and reinforced, allows the model to negotiate even conflicting information sources.

6 Acknowledgements

This research was funded by SFB 378 project "ALPHA" to the first two authors and a PhD scholarship to the last, all awarded by the German Research Foundation (DFG).

References

Altmann, G. T. M. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1):37–55.

Knoeferle, P. and Crocker, M. W. (2004a). The coordinated processing of scene and utterance: evidence from eye-tracking in depicted events. In *Proceedings of International Conference on Cognitive Science*, Allahabad, India.

Knoeferle, P. and Crocker, M. W. (2004b). Stored knowledge versus depicted events: what guides auditory sentence comprehension. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahawah, NJ: Erlbaum. 714–719.

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95:95–127.

Mayberry, M. R. and Crocker, M. W. (2004). Generating semantic graphs through self-organization. In *Proceedings of the AAAI Symposium on Compositional Connectionism in Cognitive Science*, pages 40–49, Washington, D.C.

Miikkulainen, R. (1997). Natural language processing with subsymbolic neural networks. In Browne, A., editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*, pages 120–139. Institute of Physics Publishing, Bristol, UK; Philadelphia, PA.

Noelle, D. C., Cottrell, G. W., and Wilms, F. (1997). Extreme attraction: The benefits of corner attractors. Technical Report CS97-536, Department of Computer Science and Engineering, UCSD, San Diego, CA.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

A Second Language Acquisition Model Using Example Generalization and Concept Categories

Ari Rappoport
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
arir@cs.huji.ac.il

Vera Sheinman
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
vera46@cl.cs.titech.ac.jp

Abstract

We present a computational model of acquiring a second language from example sentences. Our learning algorithms build a construction grammar language model, and generalize using form-based patterns and the learner's conceptual system. We use a unique professional language learning corpus, and show that substantial reliable learning can be achieved even though the corpus is very small. The model is applied to assisting the authoring of Japanese language learning corpora.

1 Introduction

Second Language Acquisition (SLA) is a central topic in many of the fields of activity related to human languages. SLA is studied in cognitive science and theoretical linguistics in order to gain a better understanding of our general cognitive abilities and of first language acquisition (FLA)[1]. Governments, enterprises and individuals invest heavily in foreign language learning due to business, cultural, and leisure time considerations. SLA is thus vital for both theory and practice and should be seriously examined in computational linguistics (CL), especially when considering the close relationship to FLA and the growing attention devoted to the latter by the CL community.

In this paper we present a computational model of SLA. As far as we could determine, this is the first model that simulates the learning process computationally. Learning is done from examples, with no reliance on explicit rules. The model is unique in the usage of a conceptual system by the learning algorithms. We use a unique professional language learning corpus, showing effective learning from a very small number of examples. We evaluate the model by applying it to assisting the authoring of Japanese language learning corpora.

We focus here on basic linguistic aspects of SLA, leaving other aspects to future papers. In particular, we assume that the learner possesses perfect memory and is capable of invoking the provided learning algorithms without errors.

In sections 2 and 3 we provide relevant background and discuss previous work. Our input, learner and language models are presented in section 4, and the learning algorithms in section 5. Section 6 discusses the authoring application.

2 Background

We use the term 'second language acquisition' to refer to any situation in which adults learn a new language[2]. A major concept in SLA theory [Gass01, Mitchell03] is that of **interlanguage**: when learning a new language (L2), at any given point in time the learner has a valid partial L2 language system that differs from his/her native language(s) (L1) and from the L2. The SLA process is that of progressive enhancement and refinement of interlanguage. The main trigger for interlanguage modification is when the learner notices a **gap** between interlanguage and L2 forms. In order for this to happen, the learner must be provided with **com-**

[1] Note that the F stands here for 'First', not 'Foreign'.

[2] Some SLA texts distinguish between 'second' and 'foreign' and between 'acquisition' and 'learning'. We will not make those distinctions here.

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 45–52,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

prehensible input. Our model directly supports all of these notions.

A central, debated issue in language acquisition is whether FLA mechanisms [Clark03] are available in SLA. What is clear is that SL learners already possess a mature conceptual system and are capable of explicit symbolic reasoning and abstraction. In addition, the amount of input and time available for FLA are usually orders of magnitude larger than those for SLA.

The general linguistic framework that we utilize in this paper is that of **Construction Grammar (CG)** [Goldberg95, Croft01], in which the building blocks of language are words, phrases and phrase templates that carry meanings. [Tomasello03] presents a CG theory of FLA in which children learn whole constructions as 'islands' that are gradually generalized and merged. Our SLA model is quite similar to this process.

In language education, current classroom methods use a combination of formal rules and communicative situations. Radically different is the Pimsleur method [Pimsleur05], an audio-based self-study method in which rules and explanations are kept to a minimum and most learning occurs by letting the learner infer L2 constructs from translations of contextual L1 sentences. Substantial anecdotal evidence (as manifested by learner comments and our own experience) suggests that the method is highly effective. We have used a Pimsleur corpus in our experiments. One of the goals of our model is to assist the authoring of such corpora.

3 Previous Work

There is almost no previous CL work explicitly addressing SLA. The only one of which we are aware is [Maritxalar97], which represents interlanguage levels using manually defined symbolic rules. No language model (in the CL sense) or automatic learning are provided.

Many aspects of SLA are similar to first language acquisition. Unsupervised grammar induction from corpora is a growing CL research area ([Clark01, Klein05] and references there), mostly using statistical learning of model parameters or pattern identification by distributional criteria. The resulting models are not easily presentable to humans, and do not utilize semantics.

[Edelman04] presents an elegant FLA system in which constructions and word categories are identified iteratively using a graph. [Chang04] presents an FLA system that truly supports construction grammar and is unique in its incorporation of general cognitive concepts and embodied semantics.

SLA is related to machine translation (MT), since learning how to translate is a kind of acquisition of the L2. Most relevant to us here is modern example-based machine translation (EBMT) [Somers01, Carl03], due to its explicit computation of translation templates and to the naturalness of learning from a small number of examples [Brown00, Cicekli01].

The Computer Assisted Language Learning (CALL) literature [Levy97, Chapelle01] is rich in project descriptions, and there are several commercial CALL software applications. In general, CALL applications focus on teacher, environment, memory and automatization aspects, and are thus complementary to the goals that we address here.

4 Input, Learner and Language Knowledge Models

Our ultimate goal is a comprehensive computational model of SLA that covers all aspects of the phenomenon. The present paper is a first step in that direction. Our goals here are to:

- Explore what can be learned from **example-based, small, beginner-level input corpora tailored** for SLA;
- Model a learner having a mature **conceptual system**;
- Use an L2 **language knowledge** model that supports sentence enumeration;
- Identify cognitively plausible and effective SL **learning algorithms**;
- Apply the model in **assisting the authoring** of corpora tailored for SLA.

In this section we present the first three components; the learning algorithms and the application are presented in the next two sections.

4.1 Input Model

The input potentially available for SL learners is of high variability, consisting of meta-linguistic rules, usage examples isolated for learning purposes, usage examples partially or fully understood in context, dictionary-like word definitions, free-form explanations, and more.

One of our major goals is to explore the relationship between first and second language acquisition. Methodologically, it therefore makes sense to first study input that is the most similar linguistically to that available during FLA, usage examples. As noted in section 2, a fundamental property of SLA is that learners are capable of mature understanding. Input in our model will thus consist of an ordered set of **comprehensible usage examples**, where an example is a pair of L1, L2 sentences such that the former is a translation of the latter in a certain understood context.

We focus here on modeling **beginner-level proficiency**, which is qualitatively different from native-like fluency [Gass01] and should be studied before the latter.

We are interested in **relatively small** input corpora (thousands of examples at most), because this is an essential part of SLA modeling. In addition, it is of great importance, in both theoretical and computational linguistics, to explore the limits of what can be learned from meager input.

One of the main goals of SLA modeling is to discover which input is most effective for SLA, because a substantial part of learners' input can be controlled, while their time capacity is small. We thus allow our input to be **optimized** for SLA, by containing examples that are sub-parts of other examples and whose sole purpose is to facilitate learning those (our corpus is also optimized in the sense of covering simpler constructs and words first, but this issue is orthogonal to our model). We utilize two types of such sub-examples. First, we require that new words are always presented first on their own. This is easy to achieve in controlled teaching, and is actually very frequent in FLA as well [Clark03]. In the present paper we will assume that this completely solves the task of segmenting a sentence into words, which is reasonable for a beginner level corpus where the total number of words is relatively small. Word boundaries are thus explicitly and consistently marked.

Second, the sub-example mechanism is also useful when learning a construction. For example, if the L2 sentence is 'the boy went to school' (where the L2 here is English), it could help learning algorithms if it were preceded by 'to school' or 'the boy'. Hence we do not require examples to be complete sentences.

In this paper we do not deal with phonetics or writing systems, assuming L2 speech has been consistently transcribed using a quasi-phonetic writing system. Learning L2 phonemes is certainly an important task in SLA, but most linguistic and cognitive theories view it as separable from the rest of language acquisition [Fromkin02, Medin05].

The input corpus we have used is a transcribed Pimsleur Japanese course, which fits the input specification above.

4.2 Learner Model

A major aspect of SLA is that learners already possess a mature conceptual system (CS), influenced by their life experience (including languages they know). Our learning algorithms utilize a CS model. We opted for being conservative: the model is only allowed to contain concepts that are clearly possessed by the learner before learning starts. Concepts that are particular to the L2 (e.g., 'noun gender' for English speakers learning Spanish) are not allowed. Examples for concept classes include fruits, colors, human-made objects, physical activities and emotions, as well as meta-linguistic concepts such as pronouns and prepositions. A single concept is simply represented by a prototypical English word denoting it (e.g., 'child', 'school'). A concept class is represented by the concepts it contains and is conveniently named using an English word or phrase (e.g., 'types of people', 'buildings', 'language names').

Our learners can explicitly reason about concept inter-relationships. Is-a relationships between classes are represented when they are beyond any doubt (e.g., 'buildings' and 'people' are both 'physical things').

A basic conceptual system is assumed to exist before the SLA process starts. When the input is controlled and small, as in our case, it is both methodologically valid and practical to prepare the CS manually. CS design is discussed in detail in section 6.

In the model described in the present paper we do not automatically modify the CS during the learning process; CS evolution will be addressed in future models.

As stated in section 1, in this paper we focus on linguistic SLA aspects and do not address issues such as human errors, motivation and attention. We thus assume that our learner possesses perfect memory and can invoke our learning algorithms without any mistakes.

4.3 Language Knowledge Model

We require our model to support a basic capability of a grammar: enumeration of language sentences (parsing will be reported in other papers). In addition, we provide a degree of certainty for each. The model's quality is evaluated by its applicability for learning corpora authoring assistance (section 6).

The representation is based on construction grammar (CG), explicitly storing a set of constructions and their inter-relationships. CG is ideally suited for SLA interlanguage because it enables the representation of partial knowledge: every language form, from concrete words and sentences to the most abstract constructs, counts as a construction. The generative capacity of language is obtained by allowing constructions to replace arguments. For example, (child), (the child goes to school), (<x> goes to school), (<x> <v> to school) and (X goes Z) are all constructions, where <x>, <v> denote word classes and X, Z denote other constructions.

SL learners can make explicit judgments as to their level of confidence in the grammaticality of utterances. To model this, our learning algorithms assign a **degree of certainty (DOC)** to each construction and to the possibility of it being an argument of another construction. The certainty of a sentence is a function (e.g., sum or maximum) of the DOCs present in its derivation path.

Our representation is equivalent to a graph whose nodes are constructions and whose directed, labeled arcs denote the possibility of a node filling a particular argument of another node. When the graph is a-cyclic the resulting language contains a finite number of concrete sentences, easily computed by graph traversal. This is similar to [Edelman04]; we differ in our partial support for semantics through a conceptual system (section 5) and in the notion of a degree of certainty.

5 Learning Algorithms

Our general SLA scheme is that of incremental learning – examples are given one by one, each causing an update to the model. A major goal of our model is to identify effective, cognitively plausible learning algorithms. In this section we present a concrete set of such algorithms.

Structured categorization is a major driving force in perception and other cognitive processes [Medin05]. Our learners are thus driven by the desire to form useful generalizations over the input. A generalization of two or more examples is possible when there is sufficient similarity of form and meaning between them. Hence, the basic ingredient of our learning algorithms is identifying such similarities.

To identify concrete effective learning algorithms, we have followed our own inference processes when learning a foreign language from an example-based corpus (section 6). The set of algorithms described below are the result of this study.

The basic form similarity algorithm is **Single Word Difference (SWD)**. When two examples share all but a single word, a construction is formed in which that word is replaced by an argument class containing those words. For example, given 'eigo ga wakari mas' and 'nihongo ga wakari mas', the construction (<eigo, nihongo> ga wakari mas) ('I understand English/Japanese'), containing one argument class, is created. In itself, SWD only compresses the input, so its degree of certainty is maximal. It does not create new sentences, but it organizes knowledge in a form suitable for generalization.

The basic meaning-based similarity algorithm is **Extension by Conceptual Categories (ECC)**. For an argument class W in a construction C, ECC attempts to find the smallest concept category U' that contains W', the set of concepts corresponding to the words in W. If no such U' exists, C is removed from the model. If U' was found, W is replaced by U, which contains the L2 words corresponding to the concepts in U'. When the replacement occurs, it is possible that not all such words have already been taught; when a new word is taught, we add it to all such classes U (easily implemented using the new word's translation, which is given when it is introduced.)

In the above example, the words in W are 'eigo' and 'nihongo', with corresponding concepts 'English' and 'Japanese'. Both are contained in W', the 'language names' category, so in this case U' equals W'. The language names category contains concepts for many other language names, including Korean, so it suffices to teach our learner the Japanese word for Korean ('kankokugo') at some point in the future in order to update the construction to be (<eigo, nihongo, kankokugo> ga wakari mas). This creates a new sentence 'kankokugo ga wakari mas' meaning 'I understand Korean'. An example in which U' does not equal W' is given in Table 1 by 'child' and 'car'.

L2 words might be ambiguous – several concepts might correspond to a single word. Because example semantics are not explicitly represented, our system has no way of knowing which concept is the correct one for a given construction, so it considers all possibilities. For example, the Japanese 'ni' means both 'two' and 'at/in', so when attempting to generalize a construction in which 'ni' appears in an argument class, ECC would consider both the 'numbers' and 'prepositions' concepts.

The degree of certainty assigned to the new construction by ECC is a function of the quality of the match between W and U'. The more abstract is U, the lower the certainty.

The main form-based induction algorithm is **Shared Prefix, Generated Suffix (SPGS)**. Given an example 'x y' (x, y are word sequences), if there exist (1) an example of the form 'x z', (2) an example 'x', and (3) a construction K that derives 'z' or 'y', we create the construction (x K) having a degree of certainty lower than that of K. A Shared Suffix version can be defined similarly. Requirement (2) ensures that the cut after the prefix will not be arbitrary, and assumes that the lesson author presents constituents as partial examples beforehand (as indeed is the case in our corpus).

SPGS utilizes the learner's current generative capacity. Assume input 'watashi wa biru o nomi mas' ('I drink beer'), previous inputs 'watashi wa america jin des' ('I am American'), 'watashi wa' ('as to me...') and an existing construction K = (<biru, wain> o nomi mas). SPGS would create the construction (watashi wa K), yielding the new sentence 'watashi wa wain o nomi mas' ('I drink wine').

To enable faster learning of more abstract constructions, we use generalized versions of SWD and SPGS, which allow the differing or shared elements to be a *construction* rather than a word or a word sequence.

The combined learning algorithm is: given a new example, iteratively invoke each of the above algorithms at the given order until nothing new can be learned. Our system is thus a kind of inductive programming system (see [Thompson99] for a system using inductive logic programming for semantic parsing).

Note that the above algorithms treat words as atomic units, so they can only learn morphological rules if boundaries between morphemes are marked in the corpus. They are thus more useful for languages such as Japanese than, say, for Romance or Semitic languages.

Our algorithms have been motivated by general cognitive considerations. It is possible to refine them even further, e.g. by assigning a higher certainty when the focus element is a prefix or a suffix, which are more conspicuous cognitively.

6 Results and Application to Authoring of Learning Corpora

We have experimented with our model using the Pimsleur Japanese I (for English speakers) course, which comprises 30 half-hour lessons, 1823 different examples, and about 350 words. We developed a simple set of tools to assist transcription, using an arbitrary, consistent Latin script transliteration based on how the Japanese phonemes are presented in the course, which differs at places from common transliterations (e.g., we use 'mas', not 'masu'). Word boundaries were marked during transliteration, as justified in section 4.

Example sentences from the corpus are 'nani o shi mas kaa ? / what are you going to do?', 'watashi ta chi wa koko ni i mas / we are here', 'kyo wa kaeri masen / today I am not going back', 'demo hitori de kaeri mas / but I am going to return alone', etc. Sentences are relatively short and appropriate for a beginner level learner.

Evaluating the quality of induced language models is notoriously difficult. Current FLA practice favors comparison of predicted parses with ones in human annotated corpora. We have focused on another basic task of a grammar, sentence enumeration, with the goal of showing that our model is useful for a real application, assistance for authoring of learning corpora.

The algorithm has learned 113 constructions from the 1823 examples, generating 525 new sentences. These numbers do not include constructions that are subsumed by more abstract ones (generating a superset of their sentences) or those involving number words, which would distort the count upwards. The number of *potential* new sentences is much higher: these numbers are based only on the 350 words present, organized in a rather flat CS. The constructions contain many placeholders for concepts whose words would be taught in the future, which could increase the number exponentially.

In terms of precision, 514 of the 525 sentences were judged (by humans) to be syntactically correct (53 of those were problematic semantically). Regarding recall, it is very difficult to assess formally. Our subjective impression is that the learned constructions do cover most of what a reasonable person would learn from the examples, but this is not highly informative – as indicated, the algorithms were discovered by following our own inherence processes. In any case, our algorithms have been deliberately designed to be conservative to ensure precision, which we consider more important than recall for our model and application.

There is no available standard benchmark to serve as a baseline, so we used a simpler version of our own system as a baseline. We modified ECC to not remove C in case of failure of concept match (see ECC's definition in section 5). The number of constructions generated after seeing 1300 examples is 3,954 (yielding 35,429 sentences), almost all of which are incorrect.

The applicative scenario we have in mind is the following. The corpus author initially specifies the desired target vocabulary and the desired syntactical constructs, by writing examples (the easiest interface for humans). Vocabulary is selected according to linguistic or subject (e.g., tourism, sports) considerations. The examples are fed one by one into the model (see Table 1). For a single word example, its corresponding concepts are first manually added to the CS.

The system now lists the constructions learned. For a beginner level and the highest degree of certainty, the sentences licensed by the model can be easily grasped just by looking at the constructions. The fact that our model's representations can be easily communicated to people is also an advantage from an SLA theory point of view, where 'focus on form' is a major topic [Gass01]. For advanced levels or lower certainties, viewing the sentences themselves (or a sample, when their number gets too large) might be necessary.

The author can now check the learned items for errors. There are two basic error types, errors stemming from model deficiencies and errors that human learners would make too. As an example of the former, wrong generalizations may result from discrepancies between the modeled conceptual system and that of a real person. In this case the author fixes the modeled CS. Discovering errors of the second kind is exactly the point where the model is useful. To address those, the author usually introduces new full or partial examples that would enable the learner to induce correct syntax. In extreme cases there is no other practical choice but to provide explicit linguistic explanations in order to clarify examples that are very far from the learner's current knowledge. For example, English speakers might be confused by the variability of the Japanese counting system, so it might be useful to insert an explanation of the sort 'X is usually used when counting long and thin objects, but be aware that there are exceptions'. In the scenario of Table 1, the author might eventually notice that the learner is not aware that when speaking of somebody else's child a more polite reference is in order, which can be fixed by giving examples followed by an explanation. The DOC can be used to draw the author's attention to potential problems.

Preparation of the CS is a sensitive issue in our model, because it is done manually while it is not clear at all what kind of CS people have (WordNet is sometimes criticized for being arbitrary, too fine, and omitting concepts). We were highly conservative in that only concepts that are clearly part of the conceptual system of English speakers before any exposure to Japanese were included. Our task is made easier by the fact that it is guided by words actually appearing in the corpus, whose number is not large, so that it took only about one hour to produce a reasonable CS. Example categories are names (for languages, places and people), places (park, station, toilet, hotel, restaurant, shop, etc), people (person, friend, wife, husband, girl, boy), food, drink, feelings towards something (like, need, want), self motion activities (arrive, come, return), judgments of size, numbers, etc. We also included language-related categories such as pronouns and prepositions.

7 Discussion

We have presented a computational model of second language acquisition. SLA is a central subject in linguistics theory and practice, and our main contribution is in addressing it in computational linguistics. The model's learning algorithms are unique in their usage of a conceptual system, and its generative capacity is unique in its support for degrees of certainty. The model was tested on a unique corpus.

The dominant trend in CL in the last years has been the usage of ever growing corpora. We have shown that meaningful learning can be achieved from a small corpus when the corpus has been prepared by a ‘good teacher’. Automatic identification (and ordering) of corpora subsets from which learning is effective should be a fruitful research direction for CL.

We have shown that using a simple conceptual system can greatly assist language learning algorithms. Previous FLA algorithms have in effect computed a CS simultaneously with the syntax; decoupling the two stages could be a promising direction for FLA.

The model presented here is the first computational SLA model and obviously needs to be extended to address more SLA phenomena. It is clear that the powerful notion of certainty is only used in a rudimentary manner. Future research should also address constraints (e.g. for morphology and agreement), recursion, explicit semantics (e.g. parsing into a semantic representation), word segmentation, statistics (e.g. collocations), and induction of new concept categories that result from the learned language itself (e.g. the Japanese counting system).

An especially important SLA issue is L1 transfer, which refers to the effect that the L1 has on the learning process. In this paper the only usage of the L1 part of the examples was for accessing a conceptual system. Using the L1 sentences (and the existing conceptual system) to address transfer is an interesting direction for research, in addition to using the L1 sentences for modeling sentence semantics.

Many additional important SLA issues will be addressed in future research, including memory, errors, attention, noticing, explicit learning, and motivation. We also plan additional applications, such as automatic lesson generation.

Acknowledgement. We would like to thank Dan Melamed for his comments on a related document.

References

Brown Ralf, 2000, Automated Generalization of Translation Examples, COLING ’00.

Carl Michael, Way Andy, (eds), 2003, Recent Advances in Example Based Machine Translation, Kluwer.

Chang Nancy, Gurevich Olya, 2004. Context-Driven Construction Learning. Proceedings, Cognitive Science ‘04.

Chapelle Carol, 2001. Computer Applications in SLA. Cambridge University Press. .

Cicekli Ilyas, Gu”venir Altay, 2001, Learning Translation Templates from Bilingual Translational Examples. Applied Intelligence 15:57-76, 2001.

Clark Alexander, 2001. Unsupervised Language Acquisition: Theory and Practice. PhD thesis, University of Sussex.

Clark Eve Vivienne, 2003. First Language Acquisition. Cambridge University Press.

Croft, William, 2001. Radical Construction Grammar. Oxford University Press.

Edelman Shimon, Solan Zach, Horn David, Ruppin Eytan, 2004. Bridging Computational, Formal and Psycholinguistic Approaches to Language. Proceedings, Cognitive Science ‘04.

Fromkin Victoria, Rodman Robert, Hyams Nina, 2002. An Introduction to Language, 7th ed. Harcourt.

Gass Susan M, Selinker Larry, 2001. Second Language Acquisition: an Introductory Course. 2nd ed. LEA Publishing.

Goldberg Adele, 1995. Constructions: a Construction Grammar Approach to Argument Structure. Chicago University Press.

Klein Dan, 2005. The Unsupervised Learning of Natural Language Structure. PhD Thesis, Stanford.

Levy Michael, 1997. Computer-Assisted Language Learning. Cambridge University Press.

Maritxalar Montse, Diaz de Ilarraza Arantza, Oronoz Maite, 1997. From Psycholinguistic Modelling of Interlanguage in SLA to a Computational Model. CoNLL ’97.

Medin Douglas, Ross Brian, Markman Arthur, 2005. Cognitive Psychology, 4th ed. John Wiley & Sons.

Mitchell Rosamond, Myles Florence, 2003. Second Language Learning Theories. 2nd ed. Arnold Publication.

Pimsleur 2005. www.simonsays.com, under ‘foreign language instruction’.

Somers Harold, 2001. Example-based Machine Translation. Machine Translation 14:113-158.

Thompson Cynthia, Califf Mary Elaine, Mooney Raymond, 1999. Active Learning for Natural Language Parsing and Information Extraction. ICML ’99.

Tomasello Michael, 2003. Constructing a Language: a Usage Based Theory of Language Acquisition. Harvard University Press.

	Construction	DOC	Source	Comment
1	anata / you	0	example	
2	watashi / I	0	example	
3	anata no / your	0	example	
4	watashi no / my	0	example	
5	(<anata,watashi> no)	0	SWD(3,4)	The first words of 3 and 4 are different, the rest is identical.
6	(W no), where W is <anata, watashi, Japanese word for 'we'>	-1	ECC(5)	The concept category W'={I, you, we} was found in the CS. We know how to say 'I' and 'you', but not 'we'.
7	watashi ta chi / we	0	example	
8	(W no), where W is <anata, watashi, watashi ta chi>	-2	ECC(6,7)	We were taught how to say 'we', and an empty slot for it was found in 6.
				Now we can generate a new sentence: 'watashi ta chi no', whose meaning ('our') is inferred from the meaning of construction 6.
9	chiisai / small	0	example	
10	kuruma / car	0	example	
11	chiisai kuruma / a small car	0	example	
12	watashi ta chi no kuruma / our car	0	example	
13	((W no) kuruma)	-3	SSGP (12, 11, 10, 8)	Shared Suffix Generated Prefix: (0) new example 12 = 'y x' (x: kuruma) (1) existing example 11 = 'z x' (2) existing example 10 = 'x' (3) construction K (#8) deriving 'y' learns the new construction (K x)
				Now we can generate a new sentence: 'watashi no kuruma', meaning 'my car'.
14	kodomo / child	0	example	
...	...	0	examples	Skipping a few examples...
20	((W no) kodomo)	-3	...	This construction was learned using the skipped examples.
21	((W no) <kuruma, kodomo>)	-3	SWD (13, 20)	Note that the shared element is a construction this time, not a sub-sentence.
22	((W no) P), where P is the set of Japanese words for physical things (animate or inanimate)	-4	ECC (21)	The smallest category that contains the concepts 'car' and 'child' is P'=PhysicalThings.
				Now we can generate many new sentences, meaning 'my X' where X is any Japanese word we will learn in the future denoting a physical thing.

Table 1: A learning scenario. For simplicity, the degree of certainty here is computed by adding that of the algorithm type to that of the most uncertain construction used. Note that the notation used was designed for succinct presentation and is not the optimal one for authors of learning corpora (for example, it is probably easier to visualize the sentences generated by construction #22 if it were shown as ((<watashi, anata, watashi ta chi> no) <kuruma, kodomo>).)

Item-based Constructions and the Logical Problem

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
macw@cmu.edu

Abstract

The logical problem of language is grounded on arguments from poverty of positive evidence and arguments from poverty of negative evidence. Careful analysis of child language corpora shows that, if one assumes that children learn through item-based constructions, there is an abundance of positive evidence. Arguments regarding the poverty of negative evidence can also be addressed by the mechanism of conservative item-based learning. When conservativism is abandoned, children can rely on competition, cue construction, monitoring and probabilistic identification to derive information from positive data to recover from overgeneralization.

1. The Logical Problem

Chomsky (1957, 1980) has argued that the child's acquisition of grammar is 'hopelessly underdetermined by the fragmentary evidence available.' He attributed this indeterminacy to two major sources. The first is the degenerate nature of the input. According to Chomsky, the sentences heard by the child are so full of retracing, error, and incompletion that they provide no clear indication of the possible sentences of the language. Coupled with this problem of input degeneracy is the problem of unavailability of negative evidence. According to this view, children have a hard time knowing which forms of their language are acceptable and which are unacceptable, because parents fail to provide consistent evidence regarding the ungrammaticality of unacceptable sentences. Worse still, when such evidence is provided, children appear to ignore it.

Chomsky's (1957) views about the degeneracy of the input did not stand up well to the test of time. As Newport, Gleitman & Gleitman (1977) reported, 'the speech of mothers to children is unswervingly well-formed.' More recently, Sagae, Lavie & MacWhinney (2004) examined several of the corpora in the CHILDES database and found that adult input to children can be parsed with an accuracy level parallel to that for corpora such as the Wall Street Journal database.

This evidence for well formedness of the input did not lead to the collapse of the 'argument from poverty of stimulus' (APS). However, it did place increased weight on the remaining claims regarding the absence of relevant evidence. The overall claim is that, given the absence of appropriate positive and negative evidence, no child can acquire language without guidance from a rich set of species-specific innate hypotheses. Some refer to the argument from poverty of stimulus as the 'logical problem of language acquisition (Baker, 1979), while others have called it 'Plato's Problem,' 'Chomsky's Problem,' 'Gold's Problem,' or 'Baker's Paradox.'

2. Absence of Negative Evidence

In the 1970s, generativist analyses of learnability (Wexler & Hamburger, 1973) relied primarily on an analysis presented by Gold (1967). Gold's analysis contrasted two different language-learning situations: text presentation and informant presentation. With informant presentation, the language learner can receive feedback from an infallible informant regarding the grammaticality of every candidate sentence. This corrective feedback is called 'negative evidence' and it only requires that

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 53–68,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

ungrammatical strings be clearly identified as unacceptable. Whenever the learner formulates an overly general guess about some particular linguistic structure, the informant will label the resulting structure as ungrammatical and the learner will use this information to restrict the developing grammar. Based on initial empirical results reported by Brown & Hanlon (1970), Gold argued that negative evidence is not available to the child and that language learning cannot be based on informant presentation.

Marcus (1993) has argued that the feedback that parents provide does not discriminate consistently between grammatical and ungrammatical constructions. As a result, children cannot rely on simple, overt negative evidence for recovery from overgeneralization. Although I will argue that parents provide positive evidence in a form that solves the logical problem (Bohannon *et al.*, 1990), I agree with the observation that this evidence does not constitute overt grammatical correction of the type envisioned by Gold.

3. Absence of Positive Evidence

Beginning about 1980, generative analyses of learnability began to shift away from an emphasis on the unavailability of negative evidence to arguments based on the unavailability of positive evidence. This conceptual shift led to a relative decline in attention to recovery from overgeneralization and an increase in attention to reported cases of error-free learning. For example, Chomsky's (1980) statement of the logical problem relies on the notion of error-free learning without positive evidence. The argumentation here is that, if a structure is never encountered in the input, correct use of this structure would have to indicate innate knowledge.

Researchers have claimed that the child produces error-free learning without receiving positive evidence for structures such as: structural dependency, c-command, the binding conditions, subjacency, negative polarity items, that-trace deletion, nominal compound formation, control, auxiliary phrase ordering, and the empty category principle. In each of these cases, it is necessary to assume that the underlying universal is a part of the non-parameterized core of universal grammar. If the dimension involved were parameterized, there would be a need for some form of very early parameter setting (Wexler, 1998), which could itself introduce some error. Thus, we would expect error-free learning to occur primarily for those aspects of the grammar that are completely universal and not parameterized. Parameterized features, such as subject pro-drop, could still be guided by universal grammar. However, their learning would not necessarily be error-free.

3.1. Structural dependency

The paradigm case of error-free learning is the child's obedience to the Structural Dependency condition, as outlined by Chomsky in his formal discussion with Jean Piaget (Piattelli-Palmarini, 1980). Chomsky notes that children learn early on to move the auxiliary to initial position in questions, such as 'Is the man coming?' One formulation of this rule is that it stipulates the movement of the first auxiliary to initial position. This formulation would be based on surface order, rather than structural relations. However, if children want to question the proposition given in (1), they will never produce a movement such as (2). Instead, they will always produce (3).

1. The man who is running is coming.
2. Is the man who __ running is coming?
3. Is the man who is running __ coming?'

In order to produce (3), children must be basing the movement on structure, rather than surface order. Thus, according to Chomsky, they must be innately guided to formulate rules in terms of structure.

In the theory of barriers (Chomsky, 1986), the repositioning of the auxiliary in the tree and then in surface structure involves a movement of INFL to COMP that is subject to the head movement constraint. In (2) the auxiliary would need to move around the N' of 'man' and the CP and COMP of the relative clause, but this movement would be blocked by the head movement constraint (HMC). No such barriers exist in the main clause. In addition, if the auxiliary moves as in (2), it leaves a gap that will violate the empty category principle (ECP). Chomsky's discussion with Piaget does not rely on these details. Chomsky simply argues that the child has to realize that phrasal structure is somehow involved in this process and that one cannot formulate the rule of auxiliary movement as 'move the first auxiliary to the front.'

Chomsky claims that, ‘A person might go through much or all of his life without ever having been exposed to relevant evidence, but he will nevertheless unerringly employ the structure-dependent generalization, on the first relevant occasion.’ A more general statement of this type provided by Hornstein & Lightfoot (1981) who claim that, ‘People attain knowledge of the structure of their language for which no evidence is available in the data to which they are exposed as children.’

In order to evaluate these claims empirically, we need to know when children first produce such sentences and whether they have been exposed to relevant examples in the input prior to this time. In searching for instances of relevant input as well as first uses, we should include two types of sentences. First, we want to include sentences such as (3) in which the moved verb was a copula in the relative clause, as well as sentences with auxiliaries in both positions, such as ‘Will the boy who is wearing a Yankee’s cap step forward?’ The auxiliaries do not have to be lexically identical, since Chomsky’s argument from poverty of stimulus would also apply to a child who was learning the movement rule on the basis of lexical class, as opposed to surface lexical form.

Examining the TreeBank structures for the Wall Street Journal in the Penn TreeBank, Pullum & Scholz (Pullum & Scholz, 2002) estimate that adult corpora contain up to 1% of such sentences. However, the presence of such structures in formal written English says little about their presence in the input to the language-learning child. A search by Lewis & Elman (2001) of the input to English-speaking children in the CHILDES database (MacWhinney, 2000) turned up only one case of this structure out of approximately 3 million utterances. Since CHILDES includes good sampling of target children up to age 5;0, we can safely say that positive evidence for this particular structure is seldom encountered in the language addressed to children younger than 5;0.

Because children do not produce sentences of this type themselves, it is difficult to use production data to demonstrate the presence of the constraint. Crain & Nakayama (1987) attempted to get around this problem by eliciting these forms from children directly. They asked children (3;2 to 5;11) to, ‘Ask Jabba if the boy who is watching Mickey is happy.’ Children responded with a variety of structures, none of which involved the movement of the auxiliary from the relative clause. Unfortunately, this elicitation procedure encourages children to treat the relative clause (‘the boy who is watching Mickey’) as an imitated chunk. Despite the serious methodological limitation in this particular study, it seems reasonable to believe that four-year-old children are beginning to behave in accordance with the Structural Dependency condition for sentences like (2) and (3). But does this mean that they reach this point without learning?

There is another type of sentence that provides equally useful positive evidence regarding auxiliary movement. These are wh-questions with embedded relative clauses. It turns out that there are hundreds of input sentences of this type in the CHILDES corpus. Most of these have the form of (4), but some take the form of (5).

4. Where is the dog that you like?
5. Which is the dog that is clawing at the door?

In (5) the child receives clear information demonstrating that moved auxiliaries derive from the main clause and not the relative clause. Using evidence of the type provided in (4), the child simply learns that moved auxiliaries and the wh-words that accompany them are arguments of the verb of the main clause. Sentences like (4) and (5) are highly frequent in the input to children and both types instruct the child in the same correct generalization.

Based on evidence from the main clause, the child could formulate the rule as a placement after the wh-word of the auxiliary that is conceptually related to the verb being questioned. In other words, it is an attachment to the wh-word of an argument of the main verb. This is a complex application of the process of item-based construction generation proposed in MacWhinney (1975, 1982). This formulation does not rely on barriers, ECP, HCP, INFL, COMP, or movement. It does rely on the notion of argument structure, but only as it emerges from the application of item-based constructions. Given this formulation, a few simple yes–no questions would be enough to demonstrate the pattern. When children hear ‘is the baby happy’ they can learn that the initial copula auxiliary ‘is’ takes a subject argument in the next slot and a predicate argument in the following slot. They will learn similar frames for each of the other fronted auxiliaries. When they then encounter sentences such as (11) and (12), they will further elaborate the item-based auxiliary frames to allow for positioning of the initial wh-words and for attachment of the auxiliaries to these wh-words.

One might argue that this learning scenario amounts to a restatement of Chomsky's claim, since it requires the child to pay attention to relational patterns, rather than serial order as calculated from the beginning of the sentence. However, if the substance of Chomsky's claim is that children learn to fill argument slots with compound constituents, then his analysis seems indistinguishable from that of MacWhinney (1975; 1987a).

3.2 Auxiliary phrases

Kimball (1973) presented perhaps the first example of a learnability problem based on poverty of positive evidence. He noted that children are exposed to scores of sentences with zero, one, or two auxiliaries as in (6)–(13). However, his searches of a million sentences in early machine-readable corpora located not a single instance of a structure such as (13).

6. It rains.
7. It may rain.
8. It may have rained.
9. It may be raining.
10. It has rained.
11. It has been raining.
12. It is raining.
13. It may have been raining.

Kimball argued that, despite the absence of positive data for (13), children are still able to infer its grammaticality from the data in (6) to (12). He took this as evidence that children have innate knowledge of structural compositionality. The empirical problem with Kimball's analysis is that sentences like (13) are not nearly as rare as his corpus analysis suggests. My search of the CHILDES database for the string 'might have been' located 27 instances in roughly 3 million sentences. In addition there were 24 cases of 'could have been', 15 cases of 'should have been', and 70 cases of 'would have been.' Thus, there seems to be little shortage of positive evidence for the direct learning of this pattern. Perhaps Kimball's findings to the contrary arose from focusing exclusively on 'may', since a search for 'may have been' turned up only 5 cases.

3.3 The complex-NP constraint

The complex-NP constraint blocks movement of a noun from a relative clause, as in (14) and (15).

14. *Who did John believe the man that kissed __ arrived
15. Who did John believe __ kissed his buddy?

This same constraint also blocks movement from prepositional phrases and other complex NPs, as in (16) – (18):

16. *Who did pictures of ___ surprise you?
17. *What did you see a happy ___ ?
18. *What did you stand between the wall and ___ ?

The constraint in (18) has also been treated as the coordinated-NP constraint in some accounts. Although it appears that most children obey these constraints, there are some exceptions. Wilson & Peters (1988) list these violations of the complex NP constraint from Wilson's son Seth between the ages of 3;0 and 5;0.

19. What am I cooking on a hot __ ? (stove)
20. What are we gonna look for some __ ? (houses)
21. What is this a funny __ , Dad?
22. What are we gonna push number __ ? (9)
23. Where did you pin this on my __ ? (robe)
24. What are you shaking all the __ ? (batter and milk)
25. What is this medicine for my __ ? (cold)

These seven violations all involve separation of a noun from its modifiers. Two other examples, illustrate violation of the complex-NP constraint in other environments:

26. What did I get lost at the __ , Dad?
27. What are we gonna go at Auntie and __ ?

Here, the prohibited raising involves prepositional phrases and a conjoined noun phrase. Violations of the latter type are particularly rare, but still do occur occasionally.

One might object that a theory of universal grammar should not be rejected on the basis of a few violations from a single child. However, other observers have reported similar errors. In the recordings from my sons Ross and Mark, I observed a few such violations. One occurred when my son Mark (at 5;4.4) said, 'Dad, next time when it's Indian Guides and my birthday, what do you think a picture of ___ should be on my cake?' Catherine Snow reports that at age 10;10, her son Nathaniel said, 'I have a fever, but I don't want to said, ‘I have a fever, but I don't want to be taken a temperature of.’

Most researchers would agree that violations of the complex-NP constraint are rare, but certainly not nonexistent. At the same time, the structures or meanings that might trigger these violations are also very rare, as is the input that would tell the child how to handle these structures. Given this, it seems to me that these patterns cannot reasonably be described as cases of error-free learning. Instead, we should treat them as instances of ‘low-error constructions.’ In this regard, they resemble errors such as stative progressives (‘I am knowing’) and double-object violations (‘He recommended the library the book’). As soon as we shift from error-free learning to low-error learning, we need to apply a very different form of analysis, since we now have to explain how children recover from making these overgeneralization errors, once they have produced them. This then induces us to again focus on the availability of negative evidence.

Of course, we could assume that the violation of the complex-NP constraint was a transient performance error and that, once the relevant performance factors are eliminated, the constraints of UG operate to block further wh-raising from complex noun phrases. But the important point here is that we now need to consider specific mechanisms for allowing for recovery from overgeneralization, even for what have been offered as the clearest cases of the application of universal constraints.

3.4 Binding conditions

Binding theory (Chomsky, 1981) offers three proposed universal conditions on the binding of pronouns and reflexives to referents. Sentence (28) illustrates two of the constraints. In (28), ‘he’ cannot be coreferential with ‘Bill’ because ‘Bill’ does not c-command the pronoun. At the same time, ‘himself’ must be coreferential with ‘Bill’ because it is a clausemate and does c-command ‘Bill.’

28. He said that Bill hurt himself.

When attempting to relate the logical problem to the study of the binding constraints, it is important to remember that the sentences produced or interpreted are fully grammatical. However, the interpretation in which the pronoun is coreferential with the full NP is disallowed by the binding principles. This means that, to study the imposition of the constraints, researchers must rely on comprehension studies, often with very young children.

It is well known that children often fail to apply these principles, even in carefully controlled experiments (O'Grady, 1997). Various accounts have been offered to reconcile these facts with the supposed universality of the constraint. However, one possibility that has seldom been explored is the idea that the binding conditions are learned on the basis of positive data. To illustrate the role that learning can play in this area, consider a study of long-distance movement of adjuncts by De Villiers, Roeper & Vainikka (De Villiers *et al.*, 1990). Children were divided into two age groups: 3;7 to 5;0 and 5;1 to 6;11. They were given sentences such as:

29. When did the boy say he hurt himself?
30. When did the boy say how he hurt himself?
31. Who did the boy ask what to throw?

For (29), 44% of the children gave long distance interpretations, associating ‘when’ with ‘hurt himself’, rather than ‘say.’ For (30), with a medial wh-phrase blocking a long-distance interpretation, only 6% gave long-distance responses. This shows that children were sensitive to the conditions on traces, in accord with P&P (Chomsky & Lasnik, 1993) theory. However, the fact that sensitivity to this contrast increases markedly across the two age groups indicates that children are learning this pattern. In the youngest group, children had trouble even understanding sentences with medial arguments like (31). The fact that this ability improves over time again points to learning of the possible interpretations of these structures.

Children can learn to interpret these sentences correctly by applying conservative learning principles that rely on positive data. First, they learn short-distance interpretations that attach the wh-word to the main clause. Then, when they hear sentences with medial “how” they add the additional possibility of the long-distance interpretation. However, they do this in a conservative item-based manner, limiting the new interpretation to sentences like (30) with medial “how.”

P&P theory can also provide an account of this development in terms of the setting of parameters. First, children must realize that their language allows movement, unlike Chinese. Next they must decide whether the movement can be local, as in German, or both local and distant as in English.

Finally, they must decide whether the movement is indexed by pronouns, traces, or both. However, once a parameter-setting account is detailed in a way that requires careful attention to complex cue patterns over time (Buttery, 2004; Sakas & Fodor, 2001), it can be difficult to distinguish it from a learning account. Using positive evidence, children can first learn that some movement can occur. Next, they can learn to move locally and finally they can acquire the cues to linking the moved argument to its original argument position, one by one.

3.5 Learnability or learning?

What have we learned from our examination of these four examples? First, we have seen that the application of universal constraints is not error-free. This is particularly true in the case of the binding conditions. Because the binding conditions involve parameter setting, it is perhaps not surprising that we see errors in this domain. However, we also find errors in the application of the non-parameterized constraint against raising from complex noun phrases. Only in the case of the structural dependency condition do we find no errors. However, for that structure there is also no usage at all by either parents or children, unless we consider attachment of auxiliaries to wh-words, which is quite frequent. It is possible that error-free learning exists in various other corners of syntactic, semantic, or lexical learning. But there is no evidence that error-free learning occurs in association with an absence of positive evidence. This is the crucial association that has been claimed in the literature and it is the one that we have shown to be false.

Second, for each of the four learnability problems we examined, we have seen that there are effective learning methods based on available positive evidence. This learning involves mechanisms of conservative, item-based learning followed by later generalization.

4. Multiple Solutions

Having now briefly surveyed the role of the logical problem in generative theory, we turn next to a consideration of seven factors that, operating together, allow the child to solve the logical problem. Of these seven factors, the first two are simply formal considerations that help us understand the scope of the problem. The last five are processes that can actually guide the child during acquisition.

4.1 Limiting the class of grammars

The first solution to the logical problem addresses the Gold analysis directly by showing how language can be generated from finite-state grammars (Reich, 1969). For example, Hausser (1999) has developed an efficient parser for left-associative grammars. He has shown that left-associative grammar can be expressed as a finite automaton that orders words in terms of part-of-speech categories. Because we know that finite automata can be identified from positive evidence (Hopcroft & Ullman, 1979), this means that children should be able to learn left-associative grammars directly without triggering a logical problem. Given the fact that these grammars can parse sentences in a time-linear and psycholinguistically plausible fashion, they would seem to be excellent candidates for further exploration by child language researchers.

A formal solution to the logical problem also arises in the context of the theory of categorical grammar. Kanazawa (1998) shows that a particular class of categorial grammars known as the k-valued grammars can be learned on positive data. Moreover, he shows that most of the customary versions of categorial grammar discussed in the linguistic literature can be included in this k-valued class. Shinohara (1994) and Jain, Osherson, Royer & Sharma (1999) examine still further classes of complex non-finite languages that can be learned on the basis of positive data alone. These attempts to recharacterize the nature of human language by revised formal analysis all stand as useful approaches to the logical problem. By characterizing the target language in a way that makes it learnable by children, linguists help bridge the gap between linguistic theory and child language studies.

4.2 Revised end-state criterion

The second solution to the logical problem involves resetting our notion of what it means to acquire an end-state grammar. Horning (1969) showed that, if the language identification is allowed to involve a stochastic probability of identification, rather than an absolute guarantee of no further error ever, then language can be identified on positive evidence alone. It is surprising that this solution has not received more attention, since this analysis undercuts the core logic of the logical problem, as it applies to the learning of all rule systems up to the level of context-sensitive grammars. If learning were deterministic, children would go through a series of attempts to hypothesize the 'correct' grammar for the language. Once they hit on the correct identification, they would then never abandon this end-state grammar. The fact that adults make speech errors and differ in their judgments regarding at least some syntactic structures suggests that this criterion is too strong and that the view of grammar as stochastic is more realistic.

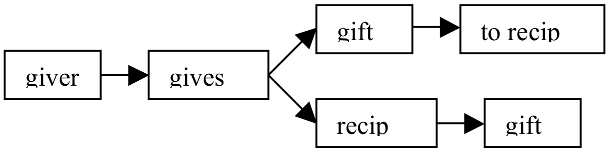
4.3 Conservative Item-based Learning

The third solution to the logical problem emphasizes the conservative nature of children's language learning. The most direct way for a language learner to solve Gold's problem is to avoid formulating overly general grammars in the first place. If the child never overgeneralizes, there is no problem of recovery from overgeneralization and no need for negative evidence or corrective feedback. Taking this basic idea one step further, let us imagine that grammars are ordered strictly in terms of their relative generative power. If this is true, then the forms generated by a grammar are a subset of the next slightly larger grammar. This is known as the Subset Principle. If the child always chooses the least powerful grammar that is consistent with the input data, then the problem of the unavailability of negative evidence disappears and learning can be based simply on positive evidence.

The Subset Principle has often been used to argue for abstract relations between grammars. For example, Fodor & Crain (1987) argue that the child learns the periphrastic dative ('give the book to John') for each new verb and only assumes that the double object construction ('give John the book') can be applied if it is attested in the input. In this particular case, the grammar with only the periphrastic is ordered as a subset of the grammar with both constructions. This follows from the principles for expansion of curly braces in GPSG.

Conservatism can control acquisition of these structures without invoking the Subset Principle. The theory of item-based acquisition (MacWhinney, 1975, 1982, 1987a; Tomasello, 2000) holds that syntactic learning is driven by the induction and combination of item-based constructions. Each item-based construction specifies a set of slots for arguments. Initially, these slots encode features that are specific to the first words encountered in this slot during comprehension. For example, the item 'more' has a slot for a following argument. If the first combinations the child picks up from comprehension are 'more cookies' and 'more milk', then this slot will initially be limited to foods. However, as the child hears 'more' used in additional combinations, the semantics of the slot filler will extend to any mass noun or plural. This learning is based entirely on generalization from positive evidence.

When learning the item-based construction for 'give', children encounter sentences such as 'Bill gives John the book.' From this, they learn the double-object construction: giver + 'give' + recipient + gift. They also learn the competing item-based construction of giver + 'give' + gift + 'to' recipient. There is no need to invoke the Subset Principle to explain this learning, since item-based constructions are inherently conservative and provide their own constraints on the form of grammars. Having acquired these two basic constructions, children can them join them into a single item-based finite automaton that operates on narrowly defined lexical categories.



Children can learn this item-based grammar fragment on the basis of simple positive data. This example uses the formalism of a finite-state automaton to annotate the use of positive data. However, in the Competition Model and other connectionist accounts, the two verb frames compete probabilistically with the outcome of the competition being determined by further cues such as focusing or topicalization.

Item-based learning involves an ongoing process of generalization for the semantic features of the arguments. During these processes of generalization, to minimize the possibility of error, the child has to be conservative in three ways:

- The child needs to formulate each syntactic combination as an item-based construction.

- Each item-based construction needs to record the exact semantic status of each positive instance of an argument in a particular grammatical configuration (MacWhinney, 1987a).
- Attempts to use the item-based construction with new arguments must be closely guided by the semantics of previously encountered positive instances.

If the child has a good memory and applies this method cautiously, overgeneralization will be minimized and there will be no need to recover from overgeneralization.

Each item-based construction is linked to a specific lexical item. This item must be a predicate. There are no item-based constructions for nouns. Predicates can have up to three arguments. Item-based constructions for verbs can also include the verbs of embedded clauses as possible arguments. Item-based constructions for prepositions and auxiliaries include both a phrase internal head (endohead) and a head for the phrase attachment (exohead). For details on the implementation of this grammatical relations model through a parser see Sagae, MacWhinney, and Lavie (2004). In section 4.6, we will see how item-based constructions are generalized to feature-based constructions in accord with the account of MacWhinney (1987a)

Conservatism also applies to non-local movement patterns. For example, Wolfe Quintero (1992) has shown that conservatism can be used to account for L2 acquisition of the wh-movement patterns. She notes that L2 learners acquire these positive contexts for wh-movement in this order:

32. What did the little girl hit __ with the block today?
33. What did the boy play with __ behind his mother?
34. What did the boy read a story about __ this morning?

Because they are proceeding conservatively, learners never produce forms such as (35):

35. *What did the boy with ___ read a story this morning?

They never hear this structure in the input and never hypothesize a grammar that includes it. As a result, they never make overgeneralizations and never attempt wh-movement in this particular context. Data from Maratsos, Kuczaj, Fox & Chalkley (1979) show that this same analysis applies to first language learners.

4.4 Competition

Conservatism is a powerful mechanism for addressing the logical problem. However, children will eventually go ‘beyond the information given’ and produce errors (Jespersen, 1922). When the child produces errors, some mechanism must force recovery. The four processes that have been proposed by emergentist theory are: competition, cue construction, monitoring, and indirect negative evidence. Each of these processes can work to correct overgeneralization. These processes are important for addressing the version of the logical problem that emphasizes the poverty of negative evidence.

The fourth solution to the problem of poverty of negative evidence relies on the mechanism of competition. Of the four mechanisms for promoting recovery from overgeneralization, competition is the most basic, general, and powerful. Psychological theories have often made reference to the notion of competition. In the area of language acquisition, MacWhinney (1978) used competition to account for the interplay between ‘rote’ and ‘analogy’ in learning morphophonology. Competition was later generalized to all levels of linguistic processing in the Competition Model. In the 1990s, specific aspects of learning in the Competition Model were formulated through both neural network theory and the ACT-R production system.

The Competition Model views overgeneralizations as arising from two types of pressures. The first pressure is the underlying analogic force that produces the overgeneralization. The second pressure is the growth in the rote episodic auditory representation of a correct form. This representation slowly grows in strength over time, as it is repeatedly strengthened through encounters with the input data. These two forces compete for the control of production. Consider the case of ‘*goed’ and ‘went’. The overgeneralization ‘goed’ is supported by analogy. It competes against the weak rote form ‘went,’ which is supported by auditory memory. As the strength of the rote auditory form for ‘went’ grows, it begins to win out in the competition against the analogic form ‘*goed’. Finally, the error is eliminated. This is the Competition Model account for recovery from overgeneralization.

The competition between two candidate forms is governed by the strength of their episodic auditory representations. In the case of the competition between ‘*goed’ and ‘went’, the overgeneralized form has little episodic auditory strength, since it is heard seldom if at all in the input. Although ‘*goed’ lacks auditory support, it has strong analogic support from the general pattern for past tense formation. In the Competition Model, analogic pressure stimulates overgeneralization and episodic auditory encoding reins it in. The analogic pressure hypothesized in this account has been described in detail in several connectionist models of morphophonological learning. The models that most closely implement the type of competition being described here are the models of MacWhinney and Leinbach (1991) for English and MacWhinney, Leinbach, Taraban & McDonald (1989) for German. In these models, there is a pressure for regularization according to the general pattern that produces forms such as ‘*goed’ and ‘*ranned’. In addition, there are weaker gang effects that lead to overgeneralizations such as ‘*stang’ for the past tense of ‘sting’.

Competition implements the notion of blocking developed first by Baker (1979) and later by Pinker (1994). Blocking is more limited than competition because it requires either strict rule-ordering or all-or-none competition. The assumption that forms are competing for the same meaning is identical to the Principle of Uniqueness postulated by Pinker (1994). Competition is also the general case of the Direct Contrast noted by Saxton (1997).

Competition goes beyond the analyses offered by Baker, Pinker, and Saxton by emphasizing the fact that the child is continually internalizing adult forms in episodic memory. Recent evidence for the power of episodic memory in infant audition (Aslin *et al.*, 1999) has underscored the power of neural mechanisms for storing linguistic input and extracting patterns from this input without conscious processing. The Competition Model assumes that children are continually storing traces of the words and phrases they hear along with tags that indicate that these phrases derive directly from adult input. When the child then comes to produce a spontaneous form, these stored forms function as an ‘oracle’ or ‘informant’, providing delayed negative evidence that corresponds (because of competition or Uniqueness) to the currently generated productive form. The ultimate source of this negative evidence is the input. Children do not use this evidence when it is initially presented. It is only later when the information is retrieved in the context of productive combinations that it provides negative evidence. This can only happen if it is clear that stored adult forms compete directly (Saxton, 1997) with productive child forms. The crucial claim of the Competition Model is that the same retrieval cues that trigger the formation of the overgeneralized productive form also trigger the retrieval of the internalized negative evidence. When these assumptions hold, there is a direct solution to the logical problem through the availability of internalized negative evidence.

To gain a better understanding of the range of phenomena that can be understood in terms of competition, let us look at examples from morphology. lexical semantics, and syntactic constructions.

4.4.1 Morphological competition

Bowerman (1987) argued that recovery from overgeneralizations such as ‘*unsqueeze’ is particularly problematic for a Competition Model account. She holds that recovery depends on processes of semantic reorganization that lie outside the scope of competition. To make her example fully concrete, let us imagine that ‘*unsqueeze’ is being used to refer to the voluntary opening of a clenched fist. Bowerman holds that there is no obvious competitor to ‘*unsqueeze.’ However, when presented with this concrete example, most native speakers will say that both ‘release’ and ‘let go’ are reasonable alternatives. The Competition Model claim is that, because there is no rote auditory support for ‘*unsqueeze,’ forms like ‘release’ or ‘let go’ will eventually compete against and eliminate this particular error.

Several semantic cues support this process of recovery. In particular, inanimate objects such as rubber balls and sponges cannot be ‘*unsqueezed’ in the same way that they can be ‘squeezed.’ Squeezing is only reversible if we focus on the action of the body part doing the squeezing, not the object being squeezed. It is possible that, at first, children do not fully appreciate these constraints on the reversibility of this particular action. However, it is equally likely that they resort to using ‘*unsqueeze’ largely because of the unavailability of more suitable competitors such as ‘release.’ An error of this type is equivalent to production of ‘falled’ when the child is having trouble remembering the correct form ‘fell.’ Or consider the competition between '*unapproved' and its acceptable competitor 'disapproved'. We might imagine that a mortgage loan application that was initially approved could then be subsequently '*unapproved.' We might have some uncertainty about the reversibility of the approval process, but the real problem is that we have not sufficiently solidified our notion of 'disapproved' in order to have it apply in this case. The flip side of this coin is that many of the child's extensional productions of reversives will end up being acceptable. For example, the child may produce 'unstick' without ever having encountered the form in the input. In this case, the form will survive. Although it will compete with 'remove', it will also receive occasional support from the input and will survive long enough for it to begin to carve out further details in the semantic scope of verbs that can be reversed with the prefix 'un-' (Li & MacWhinney, 1996).

4.4.2 Lexical competition

The same logic that can be used to account for recovery from morphological overgeneralizations can be used to account for recovery from lexical overgeneralizations. For example, a child may overgeneralize the word 'kitty' to refer to tigers and lions. The child will eventually learn the correct names for these animals and restrict the overgeneralized form. The same three forces are at work here: analogic pressure, competition, and episodic encoding. Although the child has never actually seen a 'kitty' that looks like a tiger, there are enough shared features to license the generalization. If the parent supplies the name 'tiger.' there is a new episodic encoding that then begins to compete with the analogic pressure. If no new name is supplied, the child may still begin to accumulate some negative evidence, noting that this particular use of 'kitty' is not yet confirmed in the input.

Merriman (1999) has shown how the linking of competition to a theory of attentional focusing can account for the major empirical findings in the literature on Mutual Exclusivity (the tendency to treat each object as having only one name). By treating this constraint as an emergent bias, we avoid a variety of empirical problems. Since competition is probabilistic, it only imposes a bias on learning, rather than a fixed innate constraint. The probabilistic basis for competition allows the child to deal with hierarchical category structure without having to enforce major conceptual reorganization. Competition may initially lead a child to avoid referring to a 'robin' as a 'bird,' since the form 'robin' would be a better direct match. However, sometimes 'bird' does not compete directly with 'robin.' This occurs when referring to a collection of different types of birds that may include robins, when referring to an object that cannot be clearly identified as a robin, or when making anaphoric reference to an item that was earlier mentioned as a 'robin.'

4.4.3 Syntactic frame competition

Overgeneralizations in syntax arise when a feature-based construction common to a group or 'gang' of verbs is incorrectly overextended to a new verb. This type of overextension has been analyzed in both distributed networks (Miikkulainen & Mayberry, 1999) and interactive activation networks (Elman *et al.*, 2005; MacDonald *et al.*, 1994; MacWhinney, 1987b). These networks demonstrate the same gang effects and generalizations found in networks for morphological forms (Plunkett & Marchman, 1993) and spelling correspondences (Taraban & McClelland, 1987). If a word shares a variety of semantic features with a group of other words, it will be treated syntactically as a member of the group.

Consider the example of overgeneralizations of dative movement. Verbs like 'give', 'send', and 'ship' all share a set of semantic features involving the transfer of an object through some physical medium. In this regard, they are quite close to a verb like 'deliver' and the three-argument verb group exerts strong analogic pressure on the verb 'deliver'. However, dative movement only applies to certain frequent, monosyllabic transfer verbs and not to multisyllabic, Latinate forms with a less transitive semantics such as 'deliver' or 'recommend.' When children overgeneralize and say, 'Tom delivered the library the book,' they are obeying analogic pressure from the group of transfer verbs that permit dative movement. In effect, the child has created a new argument frame for the verb 'deliver.' The first argument frame only specifies two arguments – a subject or 'giver' and an object or 'thing transferred.' The new lexical entry specifies three arguments. These two homophonous entries for 'deliver' are now in competition, just as '*goed' and 'went' were in competition. Like the entry for '*goed', the three-place entry for 'deliver' has good analogic support, but no support from episodic encoding derived from the input. Over time, it loses in its competition with the two-argument form of 'deliver' and its progressive weakening along with strengthening of the competing form leads to recovery from overgeneralization. Thus, the analysis of recovery from 'Tom delivered the library the book' is identical to the analysis of recovery from '*goed'.

4.4.4 Modeling construction strength

It may be useful to characterize the temporal course of competitive item-based learning in slightly more formal terms. To do this, we can say that a human language is generated by the application of a set of constructions that map arguments to predicates. For each item-based construction (IC), there is a correct mapping (CM) from argument to its predicates and any number of incorrect mappings (IM). The IMs receive support from analogical relations to groups of CM with similar structure. From these emerge feature-based constructions (FC). The CMs receive support from positive input, as well as analogical relations to other CMs and FCs. Each positive input increases the strength S of a matching CM by amount A. Learning of an IC occurs when the S of CM exceeds the S of each of the strongest competing IM by some additional amount. This is the dominance strength or DS.

To model language learning within this framework, we need to understand the distribution of the positive data and the sources of analogical support. From database searches and calculation of ages of learning of CM, we can estimate the number of positive input examples (P) needed to bring a CM to strength DS. For each C, if the input has included P cases by time T, we can say that a particular CM reaches DS monotonically in time T. At this point, IC is learned. Languages are learnable if their component ICs can be learned in time T. To measure learning to various levels, we can specify learning states in which there remain certain specified slow constructions (SC) that have not yet reached DS. Constructions learned by this time can be called NC or normal constructions. Thus, at time T, the degree of completion of the learning of L can be expressed as NC/NC + SC. This is a number that approaches 1.0 as T increases. The residual presence of a few SC, as well as occasional spontaneous declines in DS of CM will lead to deviations from 1.0. The study of the SCs requires a model of analogic support from FCs. In essence, the logical problem of language acquisition is then restated as the process of understanding how analogical pressures lead to learning courses that deviate from what is predicted by simple learning on positive exemplars for individual item-based constructions.

4.5 Cue construction

The fifth solution to the logical problem and the second of the solutions that promotes recovery from overgeneralization is cue construction. Most recovery from overgeneralization relies on competition. However, competition will eventually encounter limits in its ability to deal with the fine details of grammatical patterns. To illustrate these limits, consider the case of recovery from resultative overgeneralizations such as '*I untied my shoes loose'. This particular extension receives analogic support from verbs like 'shake' or 'kick' which permit 'I shook my shoes loose' or 'I kicked my shoes loose.' It appears that the child is not initially tuned in to the fine details of these semantic classifications. Bowerman (1988) has suggested that the process of recovery from overgeneralization may lead the child to construct new features to block overgeneralization. We can refer to this process as 'cue construction.'

Recovering from other resultative overgeneralizations may also require cue construction. For example, an error such as '*The gardener watered the tulips flat' can be attributed to the operation of a feature-based construction which yields three-argument verbs from 'hammer' or 'rake', as in 'The gardener raked the grass flat.' Source-goal overgeneralization can also fit into this framework. Consider, '*The maid poured the tub with water' instead of 'The maid poured water into the tub' and '*The maid filled water into the tub' instead of 'The maid filled the tub with water.' In each case, the analogic pressure from one group of words leads to the establishment of a case frame that is incorrect for a particular verb. Although this competition could be handled just by the strengthening of the correct patterns, it seems likely that the child also needs to clarify the shape of the semantic features that unify the ‘pour’ verbs and the ‘fill’ verbs.

Bowerman (personal communication) provides an even more challenging example. One can say ‘The customers drove the taxi driver crazy,’ but not ‘*The customers drove the taxi driver sad.’ The error involves an overgeneralization of the exact shape of the resultative adjective. A connectionist model of the three-argument case frame for ‘drive’ would determine not only that certain verbs license a third possible argument, but also what the exact semantic shape of that argument can be. In the case of the standard pattern for verbs like ‘drive,’ the resultant state must be terminative, rather than transient. To express this within the Competition Model context, we would need to have a competition between a confirmed three-argument form for ‘drive’ and a looser overgeneral form based only on analogic pressure. A similar competition account can be used to account for recovery from an error such as, ‘*The workers unloaded the truck empty’ which contrasts with ‘The workers loaded the truck full’. In both of these cases, analogic pressure seems weak, since examples of such errors are extremely rare in the language learning literature.

The actual modelling of these competitions in a neural network will require detailed lexical work and extensive corpus analysis. A sketch of the types of models that will be required is given in MacWhinney (1999).

4.6 Monitoring

The sixth solution to the logical problem involves children’s abilities to monitor and detect their own errors. The Competition Model holds that, over time, correct forms gain strength from encounters with positive exemplars and that this increasing strength leads them to drive out incorrect forms. If we make further assumptions about uniqueness, this strengthening of correct forms can guarantee the learnability of language. However, by itself, competition does not fully account for the dynamics of language processing in real social interactions. Consider a standard self-correction such as ‘I gived, uh, gave my friend a peach.’ Here the correct form ‘gave’ is activated in real time just after the production of the overgeneralization. MacWhinney (1978) and Elbers & Wijnen (1993) have treated this type of self-correction as involving ‘expressive monitoring’ in which the child listens to her own output, compares the correct weak rote form with the incorrect overgeneralization, and attempts to block the output of the incorrect form. One possible outcome of expressive monitoring is the strengthening of the weak rote form and weakening of the analogic forms. Exactly how this is implemented will vary from model to model

In general, retraced false starts move from incorrect forms to correct forms, indicating that the incorrect forms are produced quickly, whereas the correct rote forms take time to activate. Kawamoto (1994) has shown how a recurrent connectionist network can simulate exactly these timing asymmetries between analogic and rote retrieval. For example, Kawamoto’s model captures the experimental finding that incorrect regularized pronunciations of ‘pint’ to rhyme with ‘hint’ are produced faster than correct irregular pronunciations.

An even more powerful learning mechanism is what MacWhinney (1978) called ‘receptive monitoring.’ If the child shadows input structures closely, he will be able to pick up many discrepancies between his own productive system and the forms he hears. Berwick (1987) found that syntactic learning could arise from the attempt to extract meaning during comprehension. Whenever the child cannot parse an input sentence, the failure to parse can be used as a means of expanding the grammar. The kind of analysis through synthesis that occurs in some parsing systems can make powerful use of positive instances to establish new syntactic frames. Receptive monitoring can also be used to recover from overgeneralization. The child may monitor the form ‘went’ in the input and attempt to use his own grammar to match that input. If the result of the receptive monitoring is ‘*goed’, the child can use the mismatch to reset the weights in the analogic system to avoid future overgeneralizations.

Neural network models that rely on back-propagation assume that negative evidence is continually available for every learning trial. For this type of model to make sense, the child would have to depend heavily on both expressive and receptive monitoring. It is unlikely that these two mechanisms operate as continuously as would be required for a mechanism such as back-propagation. However, not all connectionist models rely on the availability of negative evidence. For example, Kohonen’s self-organizing feature map model

(Miikkulainen, 1993) learns linguistic patterns simply using cooccurences in the data with no reliance on negative evidence.

4.7 Indirect negative evidence

The seventh solution to the logical problem of language acquisition relies on the computation of indirect negative evidence. This computation can be illustrated with the error '*goed.' To construct indirect negative evidence in this case, children need to track the frequency of all verbs and the frequency of the past tense as marked by the regular '-ed.' Then they need to compute regular '-ed' as a percentage of all verbs. Next they need to track the frequency of the verb 'go' in all of its uses and the frequency of '*goed". To gain a bit more certainty, they should also calculate the frequency of a verb like 'jump' and the frequency of 'jumped.' With these ratios in hand, the child can then compare the ratio for 'go' with those for 'jump' or verbs in general and conclude that the attested cases of '*goed' are fewer than would be expected on the basis of evidence from verbs like 'jump.' They can then conclude that '*goed' is ungrammatical. Interestingly, they can do this without receiving overt correction.

The structures for which indirect negative evidence could provide the most useful accounts are ones that are learned rather late. These typically involve low-error constructions of the type that motivate the strong form of the logical problem. For example, children could compute indirect negative evidence that would block wh-raising from object-modifying relatives in sentences such as (37).

36. The police arrested the thieves who were carrying the loot.
37. *What did the police arrest the thieves who were carrying?
38. To do this, they would need to track the frequency of sentences such as:
39. Bill thought the thieves were carrying the loot.
40. What did Bill think the thieves were carrying?

Noting that raising from predicate complements occurs fairly frequently, children could reasonably conclude that the absence of raising from object modification position means that it is ungrammatical. Coupled with conservatism, indirect negative evidence can be a useful mechanism for avoiding overgeneralization of complex syntactic structures.

The item-based acquisition component of the Competition Model provides a framework for computing indirect negative evidence. The indirect negative evidence tracker could note that, although 'squeeze' occurs frequently in the input, '*unsqueeze' does not. This mechanism works through the juxtaposition of a form receiving episodic support ('squeeze') with a predicted inflected form ('unsqueeze').

This mechanism uses analogic pressure to predict the form '*unsqueeze.' This is the same mechanism as used in the generation of '*goed.' However, the child does not need to actually produce '*unsqueeze,' only to hypothesize its existence. This form is then tracked in the input. If it is not found, the comparison of the near-zero strength of the unconfirmed form 'unsqueeze' with the confirmed form 'squeeze' leads to the strengthening of competitors such as 'release' and blocking of any attempts to use 'unsqueeze.' Although this mechanism is plausible, it is more complicated than the basic competition mechanism and places a greater requirement on memory for tracking of non-occurrences. Since the end result of this tracking of indirect negative evidence is the same as that of the basic competition mechanism, it is reasonable to imagine that learners use this mechanism only as a fall back strategy, relying on simple competition to solve most problems requiring recovery from overgeneralization.

5. Consequences and Conclusions

This analysis suggests that we should not longer speak of language learning as being confined by the poverty of positive evidence or negative evidence. Both types of evidence are far more abundant than has been imagined. Nor should we assume that recovery from overgeneralization involves a fundamental logical problem. Recovery is supported by a set of four powerful processes (competition, cue construction, monitoring, and indirect negative evidence) that provide redundant and complementary solutions to the logical problem. In addition, we know that alternative characterizations of the nature of the target grammar can take much of the logical bit out of the logical problem. Finally, we have seen that the language addressed to children is not at all unparsable or degenerate, once a few superficial retracing structures are repaired.

We have reviewed seven solutions to the logical problem that work together to buffer the process of language acquisition. When we consider the interaction of the seven solutions in this way, we soon come to realize the pivotal role played by the item-based construction. First, the item-based construction directly enforces conservatism by requiring that each generalization of each argument frame be based on directly observable positive evidence. Second, the probabilistic competition between item-based constructions provides a meaningful way of understanding the probabilistic nature of grammar. Third, the competition between item-based constructions directly promotes recovery from overgeneralization. Fourth, the additional mechanisms of cue construction, indirect negative evidence, and monitoring serve to fine-tune the operations of competition. These processes operate particularly in those cases where uniqueness is not fully transparent or where the restriction of a general process requires additional fine-tuning of cues.

The current analysis assigns great importance to good positive data. Marcus (1993) has suggested that parents are inconsistent in their provision of negative evidence to the child. But the Competition Model assumes that it is positive data that is crucial for learning. One way in which a parent can provide crucial positive evidence is through recasting, but other methods are possible too. In various cultures and subgroups, positive evidence can be presented and focused through elicited repetition, choral recitation of stories, interaction with siblings, or games. Methods that emphasize shared attention and shared understanding can guide children toward the control of literate expression. This shared attention can arise in groups of co-wives in Central Africa just as easily as it can from isolated mother–child dyads in New England.

Recently, Hauser, Chomsky, & Fitch (2002) have argued that the core evolutionary adaptation that was required to support human language involved the introduction of a facility for recursion. The analysis in the current paper modifies and extends this claim by emphasizing the evolutionary (MacWhinney, 2005) and developmental (Tomasello, 2000) centrality of the item-based construction as the controller of recursive composition of phrases and sentences. However MacWhinney (2005) views linguistic recursion as emerging gradually from preexisting structures in spatial cognition, rather than as appearing suddenly during the Late Pleistocene. Studies of the functional neural underpinnings of recursion can go a long ways toward clarifying the details of these issues.

Acknowledgements

This work was supported by NSF Grant SBE-0354420 to the Pittsburgh Science of Learning Center.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney (Ed.), *The emergence of language* (pp. 359-380). Mahwah, NJ: Lawrence Erlbaum Associates.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry, 10*, 533-581.

Berwick, R. (1987). Parsability and learnability. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bohannon, N., MacWhinney, B., & Snow, C. (1990). No negative evidence revisited: Beyond learnability or who has to prove what to whom. *Developmental Psychology, 26*, 221-226.

Bowerman, M. (1987). Commentary. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bowerman, M. (1988). The "no negative evidence" problem. In J. Hawkins (Ed.), *Explaining language universals* (pp. 73-104). London: Blackwell.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11-54). New York: Wiley.

Buttery, P. (2004). A quantitative evaluation of naturalistic models of language acquisi**tion; the** efficiency of the Triggering Learning Algorithm compared to a Categorial Grammar Learner. *Coling 2004*, 1-8.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. (1981). *Lectures on government and binding*. Cinnaminson, NJ: Foris.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N., & Lasnik, H. (1993). The theory of principles and parameters. In J. Jacobs (Ed.), *Syntax: An international handbook of contemporary research* (pp. 1-32). Berlin: Walter de Gruyter.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63 No. 3*, 522-543.

De Villiers, J., Roeper, T., & Vainikka, A. (1990). The acquisition of long distance rules. In L. Frazier & J. De Villiers (Eds.), *Language processing and language acquisition*. Amsterdam: Kluwer.

Elbers, L., & Wijnen, F. (1993). Effort, production skill, and language learning. In C. Ferguson, L. Menn & C. Stoel-Gammon (Eds.), *Phonological development* (pp. 337-368). Timonium, MD: York.

Elman, J. L., Hare, M., & McRae, K. (2005). Cues, constraints, and competition in sentence processing. In M. Tomasello & D. Slobin (Eds.), *Beyond nature-nurture: Essays in honor of Elizabeth Bates*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fodor, J., & Crain, S. (1987). Simplicity and generality of rules in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*. Hillsdale, N.J.: Lawrence Erlbaum.

Gold, E. (1967). Language identification in the limit. *Information and Control, 10*, 447-474.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569-1579.

Hausser, R. (1999). *Foundations of computational linguistics: Man-machine communication in natural language*. Berlin: Springer.

Hopcroft, J., & Ullman, J. (1979). *Introduction to automata theory, languages, and computation*. Reading, Mass.: Addison-Wesley.

Horning, J. J. (1969). *A study of grammatical inference*: Stanford University, Computer Science Department.

Hornstein, N., & Lightfoot, D. (1981). *Explanation in linguistics: the logical problem of language acquisition*. London: Longmans.

Jain, S., Osherson, D., Royer, J., & Sharma, A. (1999). *Systems that learn*. Cambridge, MA: MIT Press.

Jespersen, O. (1922). *Language: Its nature, development, and origin*. London: George Allen and Unwin.

Kanazawa, M. (1998). *Learnable classes of categorial grammars*. Stanford, CA: CSLI Publications.

Kawamoto, A. (1994). One system or two to handle regulars and exceptions: How time-course of processing can inform this debate. In S. D. Lima, R. L. Corrigan & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 389-416). Amsterdam: John Benjamins.

Kimball, J. (1973). *The formal theory of grammar*. Englewood Cliffs, NJ: Prentice-Hall.

Lewis, J. D., & Elman, J. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *Proceedings of the 26th Annual Boston University Conference on Language Development*.

Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization, and competition: A connectionist model of the learning of English reversive prefixes. *Connection Science, 8*, 3-30.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676-703.

MacWhinney, B. (1975). Pragmatic patterns in child syntax. *Stanford Papers And Reports on Child Language Development, 10*, 153-165.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development, 43*, Whole no. 1, pp. 1-123.

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language acquisition: Vol. 1. Syntax and semantics* (pp. 73-136). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1987a). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249-308). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1987b). Toward a psycholinguistically plausible parser. In S. Thomason (Ed.), *Proceedings of the Eastern States Conference on Linguistics*. Columbus, Ohio: Ohio State University.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2005). Language evolution and human development. In D. Bjorklund & A. Pellegrini (Eds.), *Child development and evolutionary psychology*. New York: Academic.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121-157.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255-277.

Maratsos, M., Kuczaj, S. A., Fox, D. E., & Chalkley, M. A. (1979). Some empirical studies in the acquisition of transformational relations: Passives, negatives, and the past tense. In W. A. Collins (Ed.), *Children's language and communication*. Hillsdale, N.J.: Lawrence Erlbaum.

Marcus, G. (1993). Negative evidence in language acquisition. *Cognition, 46*, 53-85.

Merriman, W. (1999). Competition, attention, and young children's lexical processing. In B. MacWhinney (Ed.), *The emergence of language* (pp. 331-358). Mahwah, NJ: Lawrence Erlbaum.

Miikkulainen, R. (1993). *Subsymbolic natural language processing*. Cambridge, MA: MIT Press.

Miikkulainen, R., & Mayberry, M. R. (1999). Disambiguation and grammar as emergent soft constraints. In B. MacWhinney (Ed.), *The emergence of language* (pp. 153-176). Mahwah, NJ: Lawrence Erlbaum Associates.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and noneffects of maternal speech style. In C. Snow (Ed.), *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

O'Grady, W. (1997). *Syntactic development*. Chicago: Chicago University Press.

Piattelli-Palmarini, M. (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Cambridge MA: Harvard University Press.

Pinker, S. (1994). *The language instinct*. New York: William Morrow.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition, 49*, 21-69.

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review, 19*, 9-50.

Reich, P. (1969). The finiteness of natural language. *Language, 45*, 831-843.

Sagae, K., MacWhinney, B., & Lavie, A. (2004). Adding syntactic annotations to transcripts of parent-child dialogs. In *LREC 2004* (pp. 1815-1818). Lisbon: LREC.

Sakas, W., & Fodor, J. (2001). The structural triggers learner. In S. Bertolo (Ed.), *Language acquisition and learnability*. New York: Cambridge University Press.

Saxton, M. (1997). The Contrast Theory of negative input. *Journal of Child Language, 24*, 139-161.

Shinohara, T. (1994). Rich classes inferable from positive data: length-bounded elementary formal systems. *Information and Computation, 108*, 175-186.

Taraban, R., & McClelland, J. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language, 26*, 608-631.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences, 4*, 156-163.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*, 23-79.

Wexler, K., & Hamburger, H. (1973). On the insufficiency of surface data for the learning of transformational languages. In K. Hintikka (Ed.), *Approaches to natural language*. Dordrecht-Holland: D. Reidel.

Wilson, B., & Peters, A. M. (1988). What are you cookin' on a hot? Movement Constraints in the Speech of a Three-Year-Old Blind Child. *Language, 64, No.2*, 249-273.

Wolfe Quintero, K. (1992). Learnability and the acquisition of extraction in relative clauses and wh-questions. *Studies in Second Language Acquisition, 14*, 39-70.

Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion?

Xuân-Nga Cao Kam
PhD Program in Linguistics
The Graduate Center,
City University of New York
xkam@gc.cuny.edu

Iglika Stoyneshka
PhD Program in Linguistics
The Graduate Center,
City University of New York
idst_r@yahoo.com

Lidiya Tornyova
PhD Program in Linguistics
The Graduate Center,
City University of New York
ltornyova@gc.cuny.edu

William Gregory Sakas
PhD Programs in Computer Science and Linguistics
The Graduate Center
Department of Computer Science,
Hunter College,
City University of New York
sakas@hunter.cuny.edu

Janet Dean Fodor
PhD Program in Linguistics
The Graduate Center
City University of New York
jfodor@gc.cuny.edu

Extended Abstract

Reali & Christiansen (2003, 2004) have challenged Chomsky's most famous "poverty of stimulus" claim (Chomsky, 1980) by showing that a statistical learner which tracks transitional probabilities between adjacent words (bigrams) can correctly differentiate grammatical and ungrammatical auxiliary inversion in questions like (1) and (2):

(1) Is the little boy who is crying hurt?
(2) *Is the little boy who crying is hurt?

No examples like (1) occurred in the corpus that R&C employed, yet the grammatical form was chosen by the bigram model in 92% of the test sentence pairs. R&C conclude that no innate knowledge is necessary to guide child learners in making this discrimination, because the input evidently contains enough indirect statistical information (from other sentence types) to lead learners to the correct generalization.

R&C's data are impressive, but there is reason to doubt that they extend to other natural languages or even to other constructions in English. While replicating R&C's Experiment 1 (see Data [A]), we discovered that its success rests on 'accidental' English facts.

Six bigrams differ between the grammatical and ungrammatical versions of a sentence. (The 6 relevant bigrams for the test sentence pair (1)/(2) are shown in Table 1.) However, 86% of the correctly predicted test sentences were definitively selected by the single bigram "who is" (or "that is"), because it occurred in the corpus and none of the remaining 5 bigrams did.

Distinctive bigrams in (1)	**who is**	is crying	crying hurt
Distinctive bigrams in (2)	who crying	crying is	is hurt

Table 1. Six bigrams that differentiate *Is the little boy who is crying hurt?* from *Is the little boy who crying is hurt?* The first sentence is selected (as grammatical) solely due to the high probability of *who is*.

It can be anticipated that when there is no bigram "who/that is" in the grammatical test

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 69–71,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

sentence (e.g., in relative clauses with object-gaps, auxiliaries such as *was*, *can*, or *do-support*), the learning will be less successful. Our results confirm this prediction: object relatives like (4) and (5), where "who/that is” is not present, were poorly discriminated (see Data [B]).

(4) Is the wagon your sister is pushing red?
(5) *Is the wagon your sister pushing is red?

Results for sentences with only main verbs, requiring do-support in question-formation, like (6) and (7), were also very weak (see Data [C]).

(6) Does the boy who plays the drum want a cookie?
(7) *Does the boy who play the drum wants a cookie?

Furthermore, the powerful effect of "who/that is" in R&C’s experiment reflects no knowledge of relative clauses. It rests on the homophony of English relative pronouns with interrogative "who" and deictic "that". In R&C's training-set, the phonological/orthographic form "who" occurred as relative pronoun only 3 times, but as interrogative pronoun 44 times. R&C's analysis didn't differentiate these. (Similarly for "that": 14 relative versus 778 deictic or complementizer.)

In some languages relative pronouns are homophonous with other parts of speech (e.g., with determiners in German). We explored the possible effects of this by replacing the relative pronouns in the English corpus with “the”. Discrimination between grammatical and ungrammatical aux-inversion was poor (see Data [D]).

Many human languages lack any such superficial overlaps with relative pronouns. So unless there are other cues instead, learning can be expected to be unsuccessful in those languages too. We tested this hypothesis in two ways:

(i) We distinguished relative pronouns from their non-relative homophones in English by coding the former as “who-rel” and “that-rel” in both the corpus and the test sentences. We found a greatly reduced ability to select the grammatical aux-inversion construction (see Data [E]).

(ii) We tested verb fronting in Dutch questions, using a Dutch corpus comparable to the English corpus used by R&C (the Groningen Dutch corpus from CHILDES; approximately 21,000 utterances of child-directed speech, age 1;8 to 1;11). Due largely to verb-final word order in relative clauses, there was no one distinctive bigram that could be relied on to predict the correct choice. Performance on a set of 20 items tested so far was no better than chance (see Data [F]). Clearly, the Dutch examples provided no alternative cues for selecting the grammatical version.

Thus, the success rate in R&C’s experiment has very limited applicability. In general, bigram probability (or sentence cross-entropy, as computed in these experiments) is a poor predictor of grammaticality; e.g., the measure that prefers (1) over (2) **mis**-prefers (8) over (9):

(8) *Scared you want to the doggie.
(9) She can hear what we’re saying.

We conclude that the bigram evidence against the poverty of the stimulus for language acquisition has not been substantiated to date. It remains to be seen whether richer statistics-based inductive models will offer more robust cross-language learnability.

References

Chomsky, N. (1980). in M. Piattelli-Palmarini, (1980) *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky.* Cambridge: Harvard University Press.

Reali, F. & Christiansen, M. H. (2003). Reappraising Poverty of Stimulus Argument: A Corpus Analysis Approach. *BUCLD 28 Proceedings Supplement.*

Reali, F. & Christiansen, M. H. (2004). Structure Dependence in Language Acquisition: Uncovering the Statistical Richness of the Stimulus. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.*

Data

	% correct	% incorrect	% can't choose	# of sentence pairs tested to date	Experiment
A	87	13	0	100	Replication of R&C
B	33	15	52	100	Object-gap
C	50	50	0	50	Do-support
D	17	41	42	100	"The" replacement
E	17	39	44	100	Who-rel/That-rel
F	45	50	5	20	Dutch

Climbing the path to grammar: a maximum entropy model of subject/object learning

Felice Dell'Orletta
Dept. of Computer Science
University of Pisa
Largo Pontecorvo 3
56100 Pisa (Italy)

Alessandro Lenci
Dept. of Linguistics
University of Pisa
Via Santa Maria 36
56100 Pisa (Italy)

Simonetta Montemagni
ILC-CNR
Area della Ricerca
Via Moruzzi 1
56100 Pisa (Italy)

Vito Pirrelli
ILC-CNR
Area della Ricerca
Via Moruzzi 1
56100 Pisa (Italy)

{felice.dellorletta, alessandro.lenci, simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Abstract

In this paper, we discuss an application of Maximum Entropy to modeling the acquisition of subject and object processing in Italian. The model is able to learn from corpus data a set of experimentally and theoretically well-motivated linguistic constraints, as well as their relative salience in Italian grammar development and processing. The model is also shown to acquire robust syntactic generalizations by relying on the evidence provided by a small number of high token frequency verbs only. These results are consistent with current research focusing on the role of high frequency verbs in allowing children to converge on the most salient constraints in the grammar.

1 Introduction

Current research in language learning supports the view that developing grammatical competence involve mastering and integrating multiple, parallel, probabilistic constraints defined over different types of linguistic (and non linguistic) information (Seidenberg and MacDonald 1999, MacWhinney 2004). This is particularly clear when we focus on the core of grammatical development, namely the ability to properly identify syntactic relations. Psycholinguistic evidence shows that children learn to identify sentence subjects and direct objects by combining various types of probabilistic cues, such as word order, noun animacy, definiteness, agreement, etc. The relative prominence of each of these cues during the development of a child's syntactic competence can considerably vary cross-linguistically, mirroring their relative salience in the adult grammar system (cf. Bates *et al.* 1984).

If grammatical constraints are inherently probabilistic (Manning 2003), the path through which the child acquires adult grammar competence can be viewed as the process of building a stochastic model out of the linguistic input. Consistently with "usage-based" approaches to language acquisition (cf. Tomasello, 2000) grammatical constraints would thus emerge from language use thanks to the child's ability to keep track of statistical regularities in linguistic cues. In turn, this raises the issue of how children are able to exploit the statistical distribution of cues in the linguistic input. Various types of cross-linguistic evidence converge on the hypothesis that children are actually able to take great advantage of the highly skewed distribution of naturalistic language data. Goldberg *et al.* (2004), Matthews *et al.* (2003), Ninio (1999) among the others argue that verbs with high token frequency in the input have a facilitatory effect in allowing children to derive robust syntactic generalizations even from surprisingly minimal input. According to this model, syntactic learning is driven by a small pool of verbs occurring with the highest token frequency: they approximately correspond to so-called "light verbs" such as English *go*, *give*, *want* etc. These verbs would act as "cata-

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 72–81,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

lysts" in allowing children to converge on the most salient grammar constraints of the language they are acquiring.

In computational linguistics, *Maximum Entropy* models have proven to be robust statistical learning algorithms that perform well in a number of processing tasks (cf. Ratnaparkhi 1998). In this paper, we discuss successful application of a Maximum Entropy (*ME*) model to the processing of Italian syntactic relations. We believe that this discussion is of general interest for two basic reasons. First, the model is able to learn, from corpus data, a set of experimentally and theoretically well-motivated linguistic constraints, as well as their relative salience in the processing of Italian. This suggests that it is possible for a child to bootstrap and use this type of knowledge on the basis of a specific distribution of real language data, a conclusion that bears on the question of the role and type of innate inductive biases. Secondly, the model is also shown to acquire robust syntactic generalizations by relying on the evidence provided by a small number of high token frequency verbs only. With some qualifications, this evidence sheds light on the interaction between highly skewed language data distributions and language maturation. Robust grammar generalizations emerge on the basis of exposure to early, statistically stable and lexically underspecified evidence, thus providing a reliable backbone to children's syntactic development and later lexical organization.

In the following section we first broach the general problem of parsing subjects and objects in Italian. Section 3 describes an *ME* model of the problem. Section 4 and 5 are devoted to a detailed empirical analysis of the interaction of different feature configurations and of the interplay between verb token frequency and relevant generalizations. Conclusions are drawn in the final discussion.

2 Subjects and Objects in Italian

Children that learn how to process subjects and objects in Italian are confronted with a twofold challenge: i) the relatively free order of Italian sentence constituents and ii) the possible absence of an overt subject. The existence of a preferred Subject Verb Object (*SVO*) order in Italian main clauses does not rule out all other possible permutations of these units: in fact, they are all attested, albeit with considerable differences in distribution and degree of markedness (Bartolini *et al.* 2004).[1] Moreover, because of pro-drop, an Italian Verb Noun (*VN*) sequence can either be interpreted as a *VO* construction with subject omission (e.g. *ha dichiarato guerra* '(he) declared war') or as an instance of postverbal subject (*VS*, e.g. *ha dichiarato Giovanni* 'John declared'). Symmetrically, an *NV* sequence is potentially ambiguous between *SV* and *OV*: compare *il bambino ha mangiato* 'the child ate' with *il gelato ha mangiato* 'the ice-cream, (he) ate'.

These grammatical facts are in keeping with what we know about Italian children's parsing strategies. Bates *et al.* (1984) show that while, in English, word order is by and large the most effective cue for subject-object identification (henceforth *SOI*) both in syntactic processing and during the child's syntactic development, the same cue plays second fiddle in Italian. Bates and colleagues bring empirical evidence supporting the hypothesis that Italian children show extreme reliance on *NV* agreement and, secondly, on noun animacy, rather than word order. They conclude that the following syntactic constraints dominance hierarchy is operative in Italian: agreement > animacy > word order.

The fact that animacy can reliably be resorted to in Italian *SOI* receives indirect confirmation from corpus data. We looked at the distribution of animate subjects and objects in the Italian Syntactic Semantic Treebank (ISST, Montemagni *et al.*, 2003), a 300,000 tokens syntactically annotated corpus, including articles from contemporary Italian newspapers and periodicals covering a broad variety of topics. Subjects and objects in ISST were automatically annotated for animacy using the SIMPLE Italian computational lexicon (Lenci *et al.* 2000) as a background semantic resource. The annotation was then checked manually. Corpus analysis highlights a strong asymmetry in the distribution of animate nouns in subject and object roles: over 56.6% of ISST subjects are animate (out of a total number of 12,646), while only the 11.1% of objects are animate (out of a total number of 5,559). Such an overwhelming preference for inanimate objects in adult language data makes animacy play a very important role in *SOI*, both as a key developmental factor in the bootstrapping of the syntax-semantics mapping and as a reliable

[1] In the present paper we restrict ourselves to the case of declarative main clauses.

processing cue, consistently with psycholinguistic data.

On the other hand, the distribution of word order configurations in the same corpus shows another interesting asymmetry. *NV* sequences receive an *SV* interpretation in 95.6% of the cases, and an object interpretation in the remaining 4.4% (most of which are clitic and relative pronouns, whose preverbal position is grammatically constrained). The situation is quite different when we turn to *VN* sequences, where verb-object pairs represent 73.4% of the cases, with verb-subject pairs representing the remaining 26.6%. We infer that – at least in standard written Italian – *VS* is a much more consistently used construction than *OV*, and that the role of word order in Italian parsing is not a marginal one across the board, but rather *relative* to *VN* contexts only. In *NV* constructions there is a strong preference for a subject interpretation, and this suggests a more dynamic dominance hierarchy of Italian syntactic constraints than the one provided above.

As for agreement, it represents conclusive evidence for *SOI* only when a nominal constituent and a verb do not agree in number and/or person (as in *leggono il libro* '(they) read the book'). On the contrary, when noun and verb share the same person and number the impact of agreement on *SOI* is neutralised, as in *il bambino legge il libro* 'the child reads the book' or in *ha dichiarato il presidente* 'the president declared'. Although this ambiguity arises in specific contexts (i.e. when the verb is used in the third person singular or plural and the subject/object candidate agrees with it), it is interesting to note that in ISST: third person verb forms cover 95.6% of all finite verb forms; and, more interestingly for our present concerns, 87.9% of all *VN* and *NV* pairs involving a third person verb form contains an agreeing noun. From this we conclude that the contribution of agreement to our problem is fairly limited, as *lack* of agreement shows up only in a limited number of contexts.

All in all, corpus data lend support to the idea that in Italian *SOI* is governed by a complex interplay of probabilistic constraints of a different nature (morpho-syntactic, semantic, word order etc.). Moreover, distributional asymmetries in language data seem to provide a fairly reliable statistical basis upon which relevant probabilistic constraints can be bootstrapped and combined consistently. In the following section we shall present a ME model of how constraints and their interaction can be bootstrapped from language data.

3 A Maximum Entropy model of SOI

The Maximum Entropy (ME) framework offers a mathematically sound way to build a probabilistic model for *SOI*, which combines different linguistic cues. Given a linguistic context c and an outcome $a \in A$ that depends on c, in the ME framework the conditional probability distribution $p(a|c)$ is estimated on the basis of the assumption that no *a priori* constraints must be met other than those related to a set of features $f_j(a,c)$ of c, whose distribution is derived from the training data. It can be proven that the probability distribution p satisfying the above assumption is the one with the highest entropy, is unique and has the following exponential form (Berger *et al.* 1996):

(1) $$p(a \mid c) = \frac{1}{Z(c)} \prod_{j=1}^{k} a_j^{f_j(a,c)}$$

where $Z(c)$ is a normalization factor, $f_j(a,c)$ are the values of k features of the pair (a,c) and correspond to the linguistic cues of c that are relevant to predict the outcome a. Features are extracted from the training data and define the constraints that the probabilistic model p must satisfy. The parameters of the distribution $a_1, \ldots, a_k$ correspond to *weights* associated with the features, and determine the relevance of each feature in the overall model. In the experiments reported below feature weights have been estimated with the Generative Iterative Scaling (GIS) algorithm implemented in the AMIS software (Miyao and Tsujii 2002).

We model *SOI* as the task of predicting the correct syntactic function $f \in \{subject, object\}$ of a noun occurring in a given syntactic context s. This is equivalent to build the conditional probability distribution $p(f|s)$ of having a syntactic function f in a syntactic context s. Adopting the ME approach, the distribution p can be rewritten in the parametric form of (1), with features corresponding to the linguistic contextual cues relevant to *SOI*. The context s is a pair $<v_s, n_s>$, where v_s is the verbal head and n_s its nominal dependent in s. This notion of s departs from more traditional ways of describing an *SOI* context as a *triple* of one verb and two nouns in a certain syntactic configuration (e.g, SOV or VOS, etc.). In fact, we assume that *SOI* can be stated in terms of the more local task of establishing the grammatical function of a noun n observed in a verb-noun pair. This simplifying assumption is consistent with the claim in MacWhinney *et al.* (1984) that *SVO* word order is actually derivative from *SV* and *VO* local patterns and downplays the role of the transitive complex construction in sentence processing. Evidence in favour of this hypothesis also comes from corpus data: in ISST, there are 4,072 complete subject-verb-object-configurations, a small number if compared to the 11,584 verb tokens appearing with either a subject or an object only. Due to the comparative sparseness of canonical *SVO* constructions in Italian, it seems more reasonable to assume that children should pay a great deal of attention to both *SV* and *VO* units as cues in sentence perception (Matthews *et al.* 2004). Reconstruction of the whole lexical *SVO* pattern can accordingly be seen as the end point of an acquisition process whereby smaller units are re-analyzed as being part of more comprehensive constructions. This hypothesis is more in line with a *distributed* view of canonical constructions as derivative of more basic local positional patterns, working together to yield more complex and abstract constructions. Last but not least, assuming verb-noun pairs as the relevant context for *SOI* allows us to simultaneously model the interaction of word order variation with pro-drop in Italian.

4 Feature selection

The most important part of any ME model is the selection of the context features whose weights are to be estimated from data distributions. Our feature selection strategy is grounded on the main assumption that *features should correspond to linguistically and psycholinguistically well-motivated contextual cues*. This allows us to evaluate the probabilistic model also with respect to its ability to replicate psycholinguistic experimental results and to be consistent with linguistic generalizations.

Features are binary functions $f_{k_i,f}\ (f,s)$, which test whether a certain *cue* k_i for the function f occurs in the context s. For our ME model of *SOI*, we have selected the following types of features:

Word order tests the position of the noun wrt the verb, for instance:

(2)

$$f_{post,subj}(subj, s) = \begin{cases} 1 & if\ noun_s.pos = post \\ 0 & otherwise \end{cases}$$

Animacy tests whether the noun in s is *animate* or *inanimate* (cf. §.2). The centrality of this cue in Italian is widely supported by psycholinguistic evidence. Another source of converging evidence comes from functional and typological linguistic research. For instance, Aissen (2003) argues for the universal value of the following hierarchy representing the relative markedness of the associations between grammatical functions and animacy degrees (with each item in these scale been less marked than the elements to its right):

Animacy Markedness Hierarchy

Subj/Human > Subj/Animate > Subj/Inanimate
Obj/Inanimate > Obj/Animate > Obj/Human

Markedness hierarchies have also been interpreted as probabilistic constraints estimated form corpus data (Bresnan *et al.* 2001, Øvrelid 2004). In our ME model we have used a reduced version of the animacy markedness hierarchy in which human and animate nouns have been both subsumed under the general class *animate*.

Definiteness tests the degree of "referentiality" of the noun in a context pair s. Like for animacy, definiteness has been claimed to be associated with grammatical functions, giving rise to the following universal markedness hierarchy Aissen (2003):

Definiteness Markedness Hierarchy

Subj/Pro > Subj/Name > Subj/Def > Subj/Indef
Obj/Indef > Obj/Def > Obj/Name > Obj/Pro

According to this hierarchy, subjects with a low degree of definiteness are more marked than subjects with a high degree of definiteness (for objects the reverse pattern holds). Given the importance assigned to the definiteness markedness hierarchy in current linguistic research, we have included the definiteness cue in the ME model. It is worth remarking that, unlike animacy, in psycholinguistic experiments definiteness has not been assigned any effective role in *SOI*. This makes testing this cue in a computational model even more interesting, as a way to evaluate its effective contribution to Italian *SOI*. In our experiments, we have used a "compact" version of the definiteness scale: the definiteness cue tests whether the noun in the context pair i) is a name or a pronoun ii) has a definite article iii), has an indefinite article or iv) is a “bare” noun (i.e. with no article). It is worth saying that “bare” nouns are usually placed at the bottom end of the definiteness scale.

The three types of features above only refer to nominal cues in the context pairs. Nevertheless, specific lexical properties of the verb can also be resorted to in *SOI*. The probability for n_s to be subject or object may also depend on the specific lexical preferences of v_s. To take this lexical factor into account, we add a set of *lexical cues* to the three general feature types above. Lexical cues test animacy with respect to a specific verb v_k:

(3)

$$f_{anim,v_k,subj}(subj,s) = \begin{cases} 1 & if \quad v_k = v_s \wedge n_s = anim \\ 0 & otherwise \end{cases}$$

Lexical features provide evidence of the propensity of a given verb to have an animate (inanimate) subject or object. In fact, the verb argument structure and thematic properties may well influence the possible distribution of animate (inanimate) subjects and objects, thus overriding more general tendencies. By including lexical cues, we are thus able to test the interplay of lexical constraints with general grammatical ones.

Note that in our ME model we have not included agreement as a feature, in spite of its prominent role in Italian. The fact that agreement is often inconclusive for *SOI* (§.2) suggests that children must also acquire the ability to deal with the interplay of various concurrent constraints, none of which is singularly sufficient for the task completion this type of competence. It is exactly this area of syntactic competence that we wanted to explore with the experiments reported below (cf. MacWhinney *et al.* 1984, who similarly abstract from the dominant role of case in German *SOI*).

5 Testing feature configurations for *SOI*

The ME model for Italian *SOI* has been trained on 18,205 verb-subject/object pairs extracted from ISST. The training set was obtained by extracting all verb-subject and verb-object dependencies headed by an active verb occurring in a finite verbal construction and by excluding all cases where the position of the nominal constituent was grammatically constrained (e.g. clitic objects, relative clauses). Two different feature configurations have been used for training:

- *non-lexical feature configuration* (*NLC*), including only general features acting as global constraints: namely word order, noun animacy and noun definiteness;
- *lexical feature configuration* (*LC*), including word order, noun animacy and definiteness, and information about the verb head.

The test corpus consists of 645 verb-noun pairs extracted from contexts where agreement happens to be neutralized. Of them, 446 contained a subject (either pre- or post-verbal) and 199 contained an object (either pre- or post-verbal). The two feature configurations were evaluated by calculating the percentage of correctly assigned relations over the total number of test pairs (accuracy). As our model always assigns one syntactic relation to each test pair, accuracy equals both standard precision and recall. Finally, we have assumed a baseline score of 69%, corresponding to the result yielded by a dumb model assigning to each test pair the most frequent relation in the training corpus, i.e. subject.

5.1 Non-lexical feature configuration

Our first experiment was carried out with *NLC*. The accuracy on the test corpus is 91.5%; most errors (i.e. 96.4%) relate to the postverbal position, with 44 mistaken subjects (42 inanimate) and 9 mistaken objects (all animate). The score was confirmed by a 10-fold cross-validation on the whole training set (89.3% accuracy).

A further way to evaluate the goodness of the model is by inspecting the weights associated with feature values (Table 1).

	Subj	**Obj**
Preverbal	1,34E+00	2,10E-02
Postverbal	5,21E-01	1,47E+00
Anim	1,28E+00	3,34E-01
Inanim	8,60E-01	1,21E+00
PronName	1,22E+00	5,75E-01
DefArt	1,05E+00	1,00E+00
IndefArt	8,33E-01	1,16E+00
NoArticle	9,46E-01	1,07E+00

Table 1 – *Feature value weights in NLC*

The grey cells in Table 1 highlight the preference of each feature value for either subject or object identification: e.g. preverbal subjects are strongly preferred over preverbal objects; animate subjects are preferred over animate objects, etc. Interestingly, if we rank the *Anim* and *Inanim* values for subjects and objects, we can observe that they distribute consistently with the *Animacy Markedness Hierarchy* reported in §.4: *Subj*/*Anim* > *Subj*/*Inanim* and *Obj*/*Inanim* > *Obj*/*Anim*. Similarly, by ranking the values of the definiteness features in the *Subj* column by decreasing weight values we obtain the following ordering: *PronName* > *DefArt* > *IndefArt* > *NoArt*, which nicely fits in with the *Definiteness Markedness Hierarchy* in §.4. The so-called "markedness reversal" is observed if we focus on the values for the same features in the *Obj* column: the *PronName* feature represents the most marked option, followed by *DefArt*. The only exception is represented by the relative ordering of *IndefArt* and *NoArt* which however show very close values.

Evaluating feature salience

In order to evaluate the most reliable cues in Italian *SOI*, we have analysed the model predictions for different bundles of feature values. For each of the 16 different bundles (*b*) attested in the data, we have estimated *p*(*subj*|*b*) and *p*(*obj*|*b*):

b	*p*(*subj*|b)	*p*(*obj*|b)
Pre Anim IndefArt	0,994	0,006
Pre Anim DefArt	0,996	0,004
Pre Anim NoArt	0,995	0,005
Pre Anim PronName	0,998	0,002
Pre Inanim IndefArt	0,970	0,030
Pre Inanim DefArt	0,979	0,021
Pre Inanim NoArt	0,976	0,024
Pre Inanim PronName	0,990	0,010
Post Anim IndefArt	0,495	0,505
Post Anim DefArt	0,589	0,411
Post Anim NoArt	0,546	0,454
Post Anim PronName	0,743	0,257
Post Inanim IndefArt	0,153	0,847
Post Inanim DefArt	0,209	0,791
Post Inanim NoArt	0,182	0,818
Post Inanim PronName	0,348	0,652

Table 2 – *Subj/obj probabilities by different bundles*

The model shows a neat preference for subject when the noun is preverbal. Instead, when the noun is postverbal, function assignment is *de facto* decided by the noun animacy. Conversely, definiteness features have a much more secondary role: they can re-enforce (or weaken) the preference expressed by animacy, but they do not have the strength to determine *SOI*.

The relative salience of the different constraints acting on *SOI* can also be inferred by comparing the weights associated with individual feature values. For instance, Goldwater and Johnson (2003) show that ME can be successfully applied to learn constraint rankings in Optimality Theory, by assuming the parameter weights a_1, …, a_k as the ranking values of the constraints. The following table lists the 16 general constraints of the model by increasing weight values:

Feature	Weight
Preverbal_Obj	2,10E-02
Anim_Obj	3,34E-01
Postverbal_Subj	5,21E-01
ProName_Obj	5,75E-01
IndefArt_Subj	8,33E-01
Inanim_Subj	8,60E-01
NoArticle_Subj	9,46E-01
ArtDef_Obj	1,00E+00
DefArt_Subj	1,05E+00
NoArticle_Obj	1,07E+00
IndefArt_Obj	1,16E+00
Inanim_Obj	1,21E+00
PronName_Subj	1,22E+00
Anim_Subj	1,28E+00
Preverbal_Subj	1,34E+00
Postverbal_Obj	1,47E+00

Table 3 – *Constraint weights ranking*

The rankings in Table 3 can be used to derive the relative salience of each constraint. Lower ranked constraints correspond to more marked syntactic configurations that are then disfavoured in *SOI*. Notice that the two animacy constraints *Anim_Obj* and *Anim_Subj* are respectively placed near the bottom and the top end of the scale. Notwithstanding the low position of *Postverbal_Subj*, animacy is thus able to override the word order constraint and to produce a strong tendency to identify animate nouns as subjects, even when they appear in postverbal position (cf. Table 2 above). The constraint ranking thus confirms the interplay between animacy and word order in Italian, with the former playing a decisive role in assigning the syntactic function of postverbal nouns. On the other hand, the constraints involving noun definiteness occupy a more intermediate position in the general ranking, with very close values. This is again consistent with the less decisive role of this feature type in *SOI*, as shown above.

5.2 Lexical feature configuration

In this experiment the general features reported in Table 1 have been integrated with 4,316 verb-specific features as the ones exemplified below for the verb *dire* 'say':

*dire*_animSog	1.228213e+00
*dire*_noanimSog	7.028484e-01
*dire*_animOgg	3.645964e-01
*dire*_noanimOgg	1.321887e+00

whose associated weights show the strong preference of this verb to take animate subjects as opposed to inanimate ones as well as a preference for inanimate objects with respect to animate ones. The results achieved with *LC* on the test corpus show a significant improvement with respect to those obtained with *NLC*: the accuracy is now 95.5%, with a 4% improvement, confirmed by a 10-fold cross-validation (94.9%). Also in this case, most of the errors relate to the postverbal position (i.e. 27 out of 29), partitioned into 26 mistaken subjects and 1 mistaken object. Lexical features have been esorted to to solve most of the *NLC* errors (i.e. 34 out of 55). It is interesting to note however that lexical features can also be misleading. The *LC* results include 8 new errors, suggesting that lexical features do not always provide conclusive evidence: in fact, in 185 cases out of 645 test *VN* pairs (i.e. 28.7% of the cases) general features are preferred over lexical ones. It is also worth mentioning that the ranking of general animacy and definiteness features in *LC* actually fits in with the respective markedness hierarchies even with a better approximation than the one produced by *NLC*. Finally, the relative prominence of the different global features confirms the trend in Table 2, with word order being predominant in preverbal position and animacy playing a major role with postverbal nouns.

Both feature configurations of the ME model thus appear to comply with linguistic and psycholinguistic generalizations on *SOI*. On the linguistic side, the constraints learnt by the model are consistent with universal markedness hierarchies for grammatical relations. Secondly, the prominence of the various constraints in the model fits in well with psycholinguistic data. Consistently with the results in Bates *et al.* (1984), the model confirms the great impact of noun animacy in Italian, although in this case its key role seems to be more directly limited to the postverbal position. Conversely, the preverbal position is by itself a very strong cue for subject interpretation.

6 High frequency verbs and *SOI*

Frequency is known to play a major influence in language learning. In morphology, for example, highly frequent lexical items tend to be shorter forms, more readily accessible in the mental lexicon, independently stored as whole items (Stemberger and MacWhinney 1986) and fairly resistant to morphological overgeneralization through time, thus establishing a correlation between irregular inflected forms and frequency. Frequency has also been assigned a key role in the acquisition of syntactic constructions. In fact, Goldberg (1998) and Ninio (1999) have independently argued for the existence of a causal relation between early exposure to highly frequent light verbs and acquisition of abstract syntax-semantics mappings (constructions). Light verbs such as *want*, *put* and *go* tend to be very frequent, because they are applicable in a wider range of contexts and are learned and used at an early language maturation stage The main idea is that children's early use of these high frequency verbs is conducive to the acquisition of abstract constructional properties generalizing over particular instances.

Goldberg *et al.* (2004) motivate this hypothesis by observing that light verbs have high input frequency in the child's developmental environment and, at the same time, exhibit a low degree of semantic specialization. Hence, she argues, it takes a little abstraction step for a child to jump from actual instances of use of light verbs to the syntax-semantics association of their underlying construction. On the other hand, Ninio (1999) grounds the facilitatory role of highly frequent verbs on their being "pathbreaking" *prototypes* of the construction they instantiate, since they are the best models of the relevant combinatorial and semantic properties of their construction in a relatively undiluted fashion. However, in the case of light verb constructions, the correlation between high frequency and construction prototypicality and extension is tenuous. In fact, it is difficult to argue that frequent light verbs such as *see*, *want* or *do* exhibit a high degree of both semantic and constructional transitivity (Goldberg *et al*. 2004). This is reminiscent of the morphological behaviour of very frequent word forms in inflectional languages, as most of these forms are highly fused and show a general tendency towards irregular inflection and low morphological prototypicality. Furthermore, it is difficult to reconcile the "pathbreaking" view with the observation that frequently observed linguistic units are memorized in full, as unanalyzed wholes.

6.1 Testing the role of frequency

To address these open issues and put the alleged "pathbreaking" role of light verbs to the challenging test of a probabilistic model, we carried out a second battery of experiments to learn the general, non-lexical constraints from two training corpora of roughly equivalent size where overall type and token verb frequencies were controlled for. Both corpora are a subset of the original training set:

1. *skewed frequency corpus* (SF) – it includes 5,261 context pairs, obtained by selecting 15 verbs occurring more than 100 times in ISST (figures in parentheses give their token frequency): *essere* 'be' (2406), *avere* 'have' (708), *fare* 'do, make' (527), *dire* 'say, tell' (275), *dare* 'give' (173), *vedere* 'see' (134), *andare* 'go' (126), *sembrare* 'seem' (124), *cercare* 'try' (122), *mettere* 'put' (122), *portare* 'take' (121), *trovare* 'find' (112), *volere* 'want' (105), *lasciare* 'leave' (105), *riuscire* 'manage' (101). It is worth noticing that this set includes typical "pathbreaking" verbs;
2. *balanced frequency corpus* (BF) – this corpus includes 5,373 context pairs selected in such a way to ensure that every verb type in the original training set is attested in BF and occurs at most 6 times. For verbs occurring with a higher frequency, the pairs to be included in BF have been randomly selected.

Thus SF and BF represent two opposite training situations: SF contains few types with very high token frequencies, while BF contains a high number of verb types (i.e. 1457), with very low and uniform token frequency. These training sets resemble the structure of linguistic input used by Goldberg *et al.* (2004) for their experiments. In that case, one group of subjects was exposed to linguistic inputs in which some verbs occurred with a much higher frequency than the others; a second group of subjects was instead exposed to linguistic stimuli in which every verb occurred with roughly equal frequency. Therefore, by training our ME model on SF and BF we are able to evaluate the effective role of high token frequency verbs in driving syntactic learning.

The ME model with the general features only (i.e. *NLC*) was first trained on SF, and then tested on the 645-pair corpus in §.5, showing a 90% accuracy. The same ME model was then trained on BF, and then tested on the 645-pair corpus, scoring a 87% accuracy. The ME model trained on the skewed frequency data thus outperforms the model trained on BF in a statistically significant way (*?*^{2} = 4.97; *a*=0.05; *p-value* = 0.025).

By using a training set formed only by the verbs with the highest token frequency, the model has thus been able to acquire robust syntactic constraints for *SOI*. Once these constraints have been applied to unseen events, the model has achieved a performance comparable to the one of the general models in §.5. This is somehow even more significant if we consider that the training set was now formed by less than one-third of the pairs on which the models in §.5 were trained. Data quantity aside, the most relevant fact is that it is the way verb frequencies are distributed to determine the learning path, with a significant positive effect produced by high token frequency verbs. In the model trained on SF, feature ranking is also governed by markedness relations, and the relative prominence of the various constraints is utterly similar to the one discussed in §.5. In other terms, the results of this experiment prove that frequent verbs are actually able to act as "*catalysts*" of the syntactic acquisition process. It is possible for children to converge on the correct generalizations governing *SOI* in Italian, just by relying on the linguistic evidence provided by the most frequent verbs.

This view suggests a way out of the apparent paradox of the "pathbreaking" hypothesis: highly frequent verbs can be assumed to provide stable and consistent multiple probabilistic cues for the assignment of subject/object relations. The existence of positional patterns that occur with high token frequency may well provide a deeply entrenched and highly salient set of distributional cues that act as probabilistic constraints on constructional generalizations. We hypothesize that similar constructions of other less frequent verbs are processed, for lack of more specific overriding information, in the light of these constraints. Since processing is the result of a "conspiracy" of distributed constraints, "pathbreaking" prototypes need not be real construction exemplars but highly schematic patterns. We proved that highly frequent local positional patterns offer the right sort of constraint conspiracy.

7 General discussion

It appears that the distributional evidence of high frequency light verbs may well provide a solid cognitive anchor for sweeping perceptual generalizations on the syntax-semantics mapping. These generalizations are *local*, in that they involve positional *NV* and *VN* pairs only, and are *perceptual* as they address the issue of identifying appropriate syntactic relations by relying on perceptual features of linguistic contexts, such as position, animacy, etc. On the basis of these findings, one can reasonably argue that complex lexical constructions (in the sense of Goldberg 1998) are built upon these local patterns, by combining them in those contexts where the presence of a particular verb licenses such a combination.

The two feature configurations discussed in §.5 (i.e. *NLC* and *LC*) can thus be viewed as two successive steps along the path that leads towards the emergence of complex, lexically-driven constructions. This can actually be modeled as the incremental process of adding more and more lexical constraints to early lexicon-free generalizations (based on word order, animacy, definiteness etc.). As a result of such additional constraints, the presence of an intransitive verb may completely rule out the object interpretation of a *VN* pattern, flying in the face of a general bias towards viewing *VN* as a transitive pattern. This picture is compatible with the well-known observation that constructions are used rather conservatively by children at early stages of language maturation (Tomasello 2000). In fact, if early generalizations are mainly perceptual and local, we do not expect them to be used in production, at least until the child reaches a stage where they are combined into bigger lexically-driven constructions.

ME has proven to be a sound computational learning framework to simulate the interplay of complex probabilistic constraints in language. Our experiments confirm linguistic generalizations and psycholinguistic data for subjects and objects in Italian, while raising new interesting issues at the same time. This is the case of the role of definiteness in *SOI*. In fact, the model features neatly reproduce the definiteness markedness hierarchy, but definiteness does not appear to be really influential for subject and object processing. Various hypotheses are compatible with such results, including that definiteness is not a cue on which speakers rely for *SOI* in Italian. Another more interesting possibility is that definiteness constraints may indeed play a decisive role when the learner is asked to assign subject and object relations in the context of a more complex construction than a simple *NV* pair. Suppose that both nouns of a noun-noun-verb triple are amenable to a subject interpretation, but that one of them is a more likely subject than the other due to its being part of a definite noun phrase. Then, it is reasonable to expect that the model would select the definite noun phrase as the subject in the triple and opt for an object interpretation of the other candidate noun phrase.

As part of our future work, we plan to train the ME model on a more realistic corpus of parental input to Italian children, available in the CHILDES database (MacWhinney, 2000: http://childes.psy.cmu.edu/data/Romance/Italian). In fact, there is converging evidence that the use of particular constructions in parental speech is largely dominated by the use of each construction with one specific, highly frequent verb (e.g. *go* for the intransitive construction). The same trends noted in mother's speech to children are mirrored in children's early speech (Goldberg *et al*., 2004). Quochi (in preparation) reports a similar distributional pattern for the caused motion and intransitive motion verbs in two Italian CHILDES corpora (named "Italian-Antelmi" and "Italian-Calambrone"). If these findings are confirmed, the high accuracy of our ME model trained on the skewed frequency corpus (SF) allows us to expect an equally high accuracy when training the model on evidence from Italian parental speech.

This brings us to another related point: lack of correction/supervision in parental input. Since our ME model heavily relies on previously classified noun-verb pairs, we can legitimately wonder how easily it can be extended to simulate child language learning in an unsupervised mode. In fact, it should be appreciated that, in our experiments, comparatively little rests on supervised classification. Identification of the contextually-relevant subject is, for lack of explicit morphosyntactic clues such as agreement and diathesis, simply a matter of guessing the more likely *agent* of the action expressed by the verb on the basis of semantic and pragmatic features such as animacy, definiteness and noun position to the verb. *Mutatis mutandis*, the same holds for object identification. It is then highly likely that salient evidence for the correct subject/object classification comes to the child from direct observation of the situation described by a sentence. It is such systematic coupling of linguistic evidence from the sentence with perceptual evidence of the situation described by the sentence that can assist the child in developing interface notions such as subject, object and the like.

References

Aissen J., 2003. Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, 21: 435-483.

Bartolini R., Lenci A., Montemagni S., Pirrelli V., 2004. Hybrid constraints for robust parsing: First experiments and evaluation. *LREC2004*: 859-862.

Bates E., MacWhinney B., Caselli C., Devescovi A., Natale F., Venza V., 1984. A crosslinguistic study of the development of sentence interpretation strategies. *Child Development*, 55: 341-354.

Berger A., Della Pietra S., Della Pietra V., 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1): 39-71

Bresnan J., Dingare D., Manning C. D., 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. *Proceedings of the LFG01 Conference*, Hong Kong: 13-32.

Goldberg A. E., 1998. The emergence of the semantics of argument structure constructions. In B. MacWhinney (ed.), The *Emergence of Language*. Lawrence Erlbaum Associates, Hillsdale, N. J.: 197-212.

Goldberg A. E., Casenhiser D., Sethuraman N., 2004. Learning argument structure generalizations, *Cognitive Linguistics*.

Goldwater S., Johnson M. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In Spenader J., Eriksson A., Dahl Ö. (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. April 26-27, 2003, Stockholm University: 111-120.

Lenci A. *et al*., 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13 (4): 249-263.

Manning C. D., 2003. Probabilistic syntax. In R. Bod, J. Hay, S. Jannedy (eds), *Probabilistic Linguistics*, MIT Press, Cambridge MA: 289-341.

MacWhinney, B., 2000. The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates

MacWhinney B., Bates E., Kliegl R., 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23: 127-150.

MacWhinney B., 2004. A unified model of language acquisition. In J. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, Oxford University Press, Oxford.

Matthews D., Lieven E., Theakston A., Tomasello M., in press, The role of frequency in the acquisition of English word order, *Cognitive Development*.

Miyao Y., Tsujii J., 2002. Maximum entropy estimation for feature forests. *Proc. HLT2002*.

Montemagni S. *et al*. 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) *Treebanks. Building and Using Parsed Corpora*, Kluwer, Dordrecht: 189-210.

Ninio, A. 1999. Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, 26: 619- 653.

Øvrelid L., 2004. Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki.

Quochi, V., (in preparation). A constructional analysis of parental speech: The role of frequency and prediction in language acquisition, evidence from Italian.

Ratnaparkhi A., 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. Dissertation, University of Pennsylvania.

Seidenberg M. S., MacDonald M. C. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23(4): 569-588.

Stemberger, J., MacWhinney, B. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14:17-26.

Tomasello M., 2000. Do young children have adult syntactic competence? *Cognition*, 74: 209-253.

The Acquisition and Use of Argument Structure Constructions: A Bayesian Model

Afra Alishahi
Department of Computer Science
University of Toronto
afra@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

We present a Bayesian model for the representation, acquisition and use of argument structure constructions, which is founded on a novel view of constructions as a mapping of a syntactic form to a probability distribution over semantic features. Our computational experiments demonstrate the feasibility of learning general constructions from individual examples of verb usage, and show that the acquired knowledge generalizes to novel or low-frequency situations in language use.

1 Argument Structure Constructions

Construction grammars posit that, in addition to the idiosyncratic meanings associated with individual words or morphemes, meaning may also be directly associated with syntactic forms (e.g., Fillmore et al., 1988; Lakoff, 1987). In particular, an argument structure construction is a mapping between fundamental verb-argument relations and the syntax used to express them, as in (1) and (2) from Goldberg (1995).[1]

(1) Subj V Obj Obj$_2$ ⇔ X CAUSE Y RECEIVE Z
Ex: *Pat faxed Bill the letter.*

(2) Subj V Oblique ⇔ X MOVE Y
Ex: *The fly buzzed into the room.*

Associations between argument configurations (such as the transitive) and semantic features (such as causation) capture important linguistic regularities, and appear to play a role in both child language acquisition (Gleitman, 1990; Naigles, 1990; Fisher, 2002) and adult sentence interpretation (Bencini and Goldberg, 2000). An established form-meaning mapping may even impose an "unusual" meaning when a verb is used in a manner that is not typical for it. For example, in (2) above, using the verb *buzzed* in a construction with a path argument induces a semantics of movement as well as the standard sound emission sense for the verb (Goldberg, 1995).

A theory of grammar that includes argument structure constructions (henceforth, simply "constructions") as an organizing component of predicate knowledge must address a number of questions concerning the nature, acquisition and use of such constructions, namely:

- Precisely what constitutes the form-meaning mapping of a construction?
- How are general constructions learned from specific usages of verbs?
- What role do constructions play in language interpretation and production?

We have developed a Bayesian model of language acquisition and processing that has been shown to mimic children's behaviour in forming word order generalizations, and in recovering from overgeneralizations without negative evidence (Alishahi and Stevenson, 2005). In this paper, we describe how the model enables a new view on the nature of constructions, thus providing an answer to the first question above which leads to interesting consequences for the others. Specifically, each construction is not simply a form-meaning pair (as in (1) and (2) above),

[1] Such constructions serve a similar purpose to linking rules (e.g., Pinker, 1989) or the event structure templates of Rappaport Hovav and Levin (1998).

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 82–90,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

but rather maps a form to a probability distribution over the associated elements of meaning. The probabilistic nature of constructions in the model enables it to capture both statistical effects in language learning, and adaptability in language use.

Statistical patterns are widely accepted to play a role in language acquisition, and work on construction learning has taken a usage-based approach (e.g., Goldberg, 1995; Langacker, 1999; Tomasello, 2000). Our model elaborates on this view in the context of our assumption of a probabilistic form-meaning mapping. Constructions arise through an unsupervised Bayesian categorization process that groups verb usages according to probabilities over the properties of the verb and its arguments. Each group forms a construction in which the semantic primitives occurring most frequently across the group have the highest probabilistic association with the syntactic form. We demonstrate in computational experiments that such primitives are typically the more general semantic properties, modeling the ability of a child to capture argument structure regularities by inducing general constructions from individual usages.

The probabilistic nature of our view of constructions also influences the properties of language use, which we formulate as a Bayesian prediction problem. For example, in production, the model predicts the syntax to express an intended semantics, while in comprehension, it predicts (some of) the semantics of an observed utterance. Constructions enable the model to generalize observed patterns of association to new or low-frequency situations. This property underlies our further experimental results that mimic children's ability to infer the basic semantics of a novel verb from its usage. This property extends to the use of a known verb in an unusual pattern (cf. (2) above): we additionally show that the model can associate a novel construction with a verb while avoiding inappropriate overgeneralizations.

Scene-Utterance Input Pair
$\text{PUT}_{[cause, move]}(\text{MOM}_{\langle agent \rangle}, \text{TOYS}_{\langle theme \rangle}, \text{IN}_{[]}(\text{BOXES}_{\langle goal \rangle}))$ *Mom put toys in boxes.*

Extracted Argument Structure Frame		
head verb		*put*
verb sem. primitives		⟨*cause, move*⟩
args	roles	⟨*agent, theme, goal*⟩
	categories[a]	⟨*human, concrete, dest-pred*⟩
syntactic pattern		arg1 *verb* arg2 arg3

[a] Extracted from a representation of the child's ontology.

Figure 1: An input pair and extracted frame.

2 Probabilistic Constructions

2.1 Argument Structure Frames

In our view, a construction is a group of individual verb usages, the latter of which we represent as argument structure frames. Each frame records the syntactic pattern of a verb usage, along with the semantic properties of the verb and its arguments in that pattern. Our model extracts the features of a frame from an input observation in the form of a scene-utterance pair: the perceived utterance (what the child hears), and a logical form representation of the relevant aspect of the observed scene (the semantics described by the utterance).[2] Figure 1 shows a sample input pair and the extracted frame.

2.2 Constructions as Groups of Frames

Each construction is a group of extracted frames which share a common syntax—i.e., a particular syntactic configuration of arguments.[3] Because the semantic properties of such usages may vary in their particulars, elements of meaning in a construction are probabilistically associated with the syntactic form. For example, usages such as *Jay got a tower* and *Kay made a tower* may yield frames that form a (transitive) construction. While the frames share the verb semantic primitive *act*, they differ in others (*possess* for the former, and *become* for the latter). If this observation holds across a number

[2] We assume that the (non-trivial) task of picking out the utterance semantics from the full scene representation has been performed (as in Siskind, 1996, for example).

[3] Though note that not all frames with the same syntax necessarily form a single construction.

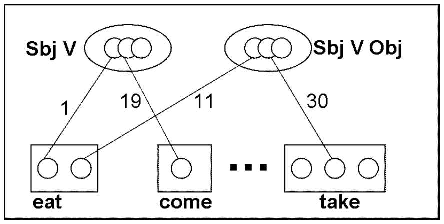


Figure 2: A portion of the lexicon showing 2 constructions. Circles represent frames.

of usages that exhibit this form, then we would find a higher probability for the primitive *act* given this construction than for the others. In this way, constructions probabilistically generalize the semantics of a set of frames.

Each verb with an observed usage that participates in a construction has a link to that construction in the lexicon. The links are weighted by the frequency with which the verb has been seen in a compatible frame, capturing the statistical usage pattern of the verb. These frequencies are also used in the calculation of the probabilities of association between a construction and the features occurring in its observed frames. Moreover, the sum of its incoming link frequencies contributes to the overall probability of a construction. Figure 2 illustrates a portion of the acquired lexicon; in the next section, we describe how the link between a frame and a construction is established.

3 Acquisition of Constructions

3.1 Overview

Every new frame is input to an incremental Bayesian clustering process that groups the new frame together with an existing group of frames—a construction—that probabilistically has the most similar properties to it. If none of the existing constructions has sufficiently high probability, then an entry for a new construction is created, containing only the new frame.

The probability of each construction for the frame is determined by both syntactic and semantic features. Currently, a construction with a different syntactic pattern from that of the frame, or a different set of argument roles, would have a very low probability. The probability of the semantic primitives of the verb, as well as the semantic categories of its arguments, is determined by how frequently those of the frame occur across the frames of the candidate construction. Because these probabilities may be calculated over partial information, we can simulate learning in the face of an incomplete frame.

3.2 Details of the Learner

The Bayesian approach we use is an adaptation of a model of human categorization proposed by Anderson (1991), which incrementally groups perceived items (in our case, frames) into categories of items with similar features (in our case, constructions).[4] It is important to note that the categories (i.e., constructions) are not predefined, but rather are determined by the similarity patterns over observed frames.

Grouping a frame F with other frames participating in construction k is formalized as finding the k with the maximum probability given F:

$$\textbf{BestConstruction}(F) = \underset{k}{\operatorname{argmax}}\ P(k|F) \quad (1)$$

where k ranges over the indices of all constructions, including an index of 0 to represent recognition of a new construction. Using Bayes rule, and dropping $P(F)$ which is constant for all k:

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \approx P(k)P(F|k) \quad (2)$$

The prior probability of k is given by:[5]

$$P(k) = \frac{n_k}{n+1} \quad (3)$$

where n is the total number of observed frames; n_k is the number of frames participating in construction k, for $k > 0$; and $n_0 = 1$. Thus, the prior for an existing construction is proportional to the frequency of its frames, and that of a new construction is inversely proportional to the number of observed frames overall. The prior probability estimation for each construction follows the intuition that it is more probable for

[4] In Alishahi and Stevenson (2005), we referred to these categories as 'classes'; we use the terms 'constructions' and 'classes' interchangeably to refer to a group of similar frames.

[5] This is the formula used by Anderson (1991) with his "coupling probability" set to the mid value of 0.5.

a frame to come from a more entrenched construction (i.e., one with more frames), and that as the number of the observed frames increases, the probability that a frame comes from a new construction decreases.

The probability of a frame F is expressed in terms of the individual probabilities of its features (shown above in Figure 1). To make the calculation feasible, we assume that these features are independent; thus, the conditional probability of a frame F is the product of the conditional probabilities of its features:

$$P(F|k) = \prod_{i \in FrameFeatures} P_i(j|k) \tag{4}$$

where j is the value of the i^{th} feature of F, and $P_i(j|k)$ is the probability of displaying value j on feature i within construction k. This probability is estimated using a smoothed maximum likelihood formulation, reflecting the emphasis on usage statistics in child language acquisition.

4 Language Use as Prediction

4.1 Overview

We formulate language use (production and comprehension) as a prediction process, in which missing features in a frame are set to the most probable values given the available features. If a usage of a verb is sufficiently complete and frequent, then it will have a strong influence on the determination of unknown features. On the other hand, constructions play an important role in the face of missing or low-confidence verb-based information, because a prediction based on a construction generalizes over all its frames. This enables the model to produce or understand a verb in a novel (for that verb) syntactic pattern, as long as semantically similar verb usages have been observed.

4.2 Details of the Prediction Process

To integrate verb-based and construction-based knowledge, we must extend the prediction aspect of the model of Anderson (1991). We begin with his prediction formula:

$$\begin{aligned} \mathbf{BestValue}_i(F) &= \underset{j}{\operatorname{argmax}}\ P_i(j|F) \\ &= \underset{j}{\operatorname{argmax}} \sum_k P_i(j|k)P(k|F) \end{aligned} \tag{5}$$

where F is a partial frame, i is a missing feature, j ranges over possible values of i, and k ranges over all categories. Intuitively, the model generalizes over the items within a category to predict the most probable value for the missing feature across those items ($P_i(j|k)$); this is then weighted by the probability of the category given the partial frame ($P(k|F)$).

However, the structure of our acquired knowledge is more complex than that of Anderson's model. Specifically, we have two groups of items over which we might generalize: in addition to the groupings of frames into constructions (which would be indicated by the formula above), we also have the groups of frames associated with each verb.

Given the importance of verb-based knowledge in language acquisition (Tomasello, 2000), we begin by focusing only on the frames associated with the verb v in frame F:

$$P_i(j|F) = \sum_{k_v \in C(v)} P_i(j|k_v)P(k_v|F) \tag{6}$$

where $C(v)$ is the set of constructions linked to by the frames of verb v. This formulation generalizes over the constructions associated with a verb, but ignores all other constructions. This unnecessarily restricts the model when a partial frame for a verb does not match well with any frame previously seen *for that verb*, but may be compatible with another learned construction.

To allow more general knowledge of constructions to influence the prediction process, we find the best construction for a partial frame F during prediction as if it were a newly observed frame to be learned. We apply our Bayesian learner to determine the most compatible construction k_F for the partial frame F using eqn. (1), and temporarily insert the frame into the lexical entry for v with a frequency of 1 on the link to k_F.[6] This ensures that the overall best construction is taken into consideration, along with the constructions associated with the verb, in predicting values for a partial frame.

$P(k_v|F)$ in eqn. (6) is rewritten using Bayes rule and dropping the $P(F)$ term (cf. eqn. (2)):

$$P(k_v|F) \approx P(k_v)P(F|k_v) \tag{7}$$

$P(F|k_v)$ is determined as in eqn. (4), using a uniform probability distribution over the possible values of the missing feature. In calculating $P(k_v)$, the frequency of each construction (its number of frames) is weighted by the frequency of v's frame which links to it, balancing the overall likelihood of the construction with the likelihood that it is a construction for v:

$$w_{k_v} = \frac{n_{k_v}}{\sum_{k_{v'} \in C(v)} n_{k_{v'}}} \times \mathit{freq}(v, k_v) \tag{8}$$

where n_{k_v} is the frequency of the construction k_v, $C(v)$ is the set of constructions linked to by the frames of verb v, $\mathit{freq}(v, k_v)$ is the frequency of v's frame which links to construction k_v, and $\mathit{freq}(v, k_F) = 1$. The prior probability $P(k_v)$ is then calculated by normalizing the weight w_{k_v} from eqn. (8):

$$P(k_v) = \frac{w_{k_v}}{\sum_{k_{v'} \in C(v)} w_{k_{v'}}} \tag{9}$$

A particularly interesting situation arises when the best overall construction, k_F, is previously unseen for v. The calculation for $P(k_v)$ entails that k_F generally has less influence the more often the verb has been seen overall; that is, greater weight from v's observed frames to their constructions will outweigh the influence of the single partial frame to its "new" construction. This factor is responsible for recovery from overgeneralization errors (Alishahi and Stevenson, 2005). However, this effect is modulated by the other factor, $P(F|k_v)$. If the partial frame F is more compatible with the "new" construction than with an existing construction for v, the use of k_F may increase in probability even for a frequent verb. This can be the situation in productive generalizations (the use of "unusual" constructions), as we demonstrate in our experimental results below.

5 Experimental Materials

For our computational experiments, we automatically create input corpora—sequences of scene-utterance pairs—using early-acquired verbs in distributional patterns of child-directed speech. The input-generation lexicon includes the 13 most frequent verbs in mother's speech across three children in CHILDES (age 1;6 to 5;1), along with their argument structure frames and associated frequencies, determined through manual analysis of the children's conversations. Additional words in the lexicon include prepositions used in the same conversations, as well as a small number of nouns.

We enumerated a set of 9 primitives for describing the coarse-level semantics of a verb (*act*, *cause*, *move*, *become*, etc.), along with features as needed to capture finer-grained meaning distinctions (such as *consume* and *rest*). The semantic categories of the nouns were selected to reflect a simplified early ontology of a child. We assume that at the stage of acquisition being modeled, these features, along with the semantic roles of the arguments, can be largely determined by the child from the observed scene. The corresponding syntactic forms are simplified to indicate the order of arguments, with words used only in their root form.

For each simulation, a random sequence of input pairs is produced from the input-generation lexicon, using the frequencies to determine the probabilities of selecting a particular verb and argument structure for each input. Arguments which are predicates (such as prepositional phrases) are constructed recursively. To simulate noise, every third input pair in every generated corpus has one of its features randomly removed. During a simulation, each missing feature is replaced with the most probable value predicted at that point in learning, corresponding to a child learning from her own inferred knowledge. The resulting input data is noisy, especially in the initial stages of learning.

[6] Note that k_F may or may not be a construction already linked to by a frame of v.

6 Experimental Results

We report computational experiments that demonstrate the ability of our model to learn constructions representing the knowledge of argument structure regularities, and to generalize this knowledge to novel situations in language use. All reported results are averaged over 10 simulations using different randomly generated input corpora; all simulations use the same system parameter settings.

We first show how general constructions emerge as the model is exposed to verb usages over time. We then turn to use of the constructions in novel situations. We demonstrate that the model induces the coarse meaning of an unknown verb based on its syntactic usage, and deals appropriately with cases where a verb appears in a usage that is unusual for it.

6.1 The Emergence of Constructions

Goldberg (1995) suggests that an argument structure construction is a grouping of similar verbs around a *light verb*, a semantically simple and highly frequent verb such as *go*, *make*, or *give*. The syntactic form and semantic properties of the core light verb determine the form-meaning mapping of the construction.

We propose an alternative view, in which constructions arise regardless of the generality of any particular verb's semantic properties. (Though constructions may be more likely to form around light verbs due to their frequency.) Even if none of its individual verb usages is semantically simple, the more-basic verb semantic primitives associated with a construction will, over time, increase in probability. While an increasing number of semantic primitives may be associated with a construction given exposure to a greater variety of verbs, the more-basic semantic primitives will become entrenched because they will be observed across many more frames.

Simple intransitive and transitive constructions emerge consistently in all our simulations. Table 1 shows the probabilistic association of verb semantic primitives with the intransitive usage ("Subj V"), after 100 and after 1000 input pairs (averaged over 10 simulations). The probabilities indicate that the construction has a very strong bias toward the general primitives *act* and *move*. Although a few primitives that are variously associated with some intransitive verb usages increase slightly in probability (*manner*, *consume*, and *rest*), the remaining primitives drop off to negligible values. An examination of the transitive construction shows a similar pattern, with primitives *act*, *possess*, and *cause* strongly entrenched after 1000 inputs.

Verb Sem.	# Input Pairs	
Primitives	100	1000
act	.47	.50
move	.37	.34
manner	.09	.11
consume	.01	.03
rest	.01	.01
possess	.01	10^{-4}
cause	.01	10^{-4}
become	.01	10^{-4}
change-state	.01	10^{-4}
perceive	.01	10^{-4}
contact	.01	10^{-4}

Table 1: Probabilities of semantic primitives associated with an intransitive construction.

6.2 Inferring Verb Meaning

A number of experiments have indicated that children use the evidence of syntactic form to infer general semantic properties of a novel verb (e.g., Naigles, 1990; Fisher, 2002), a phenomenon known as syntactic bootstrapping (Gleitman, 1990). For example, Naigles (1990) showed that children who heard an intransitive utterance with a novel verb (U_I:*The bunny and duck are blicking*) were more likely to look at a picture of two characters independently performing an action, while those who heard a similar transitive form (U_T: *The bunny is blicking the duck*) were more likely to look at a picture of one character (*the bunny*) performing an action on the other, when told to "find" the novel action in the pair of scenes. The children have learned a reliable association between a syntactic form (such as the transitive) and a coarse semantics for the expressed event (i.e., one participant causally affecting another), and are able to determine the scene that is more compatible with an utterance according to this acquired knowledge.

We demonstrate this ability in our model as follows. We create scene representations corresponding to the pictures shown to the children:

$$S_{CA}: \ \text{BLICK}_{[cause,act]}(\text{DUCK}_{\langle agent\rangle}, \text{BUNNY}_{\langle theme\rangle})$$

$$S_{A}: \ \text{BLICK}_{[act]}(\text{AND}_{[]}(\text{DUCK}, \text{BUNNY})_{\langle agent\rangle})$$

In one condition, analogous to the child hearing the transitive form, each of the above scenes is combined with the transitive utterance, U_T, to form two input pairs, S_{CA}-U_T and S_A-U_T. The former input pair corresponds to the appropriate selection of the scene to go along with the perceived utterance, and the latter to the inappropriate selection. (We form analogous pairings with the intransitive form, U_I.) We then (separately) input each pair to our model to have it extract a corresponding frame F and determine the best construction k for it. The model records the value from eqn. (1), $P(k|F)$, as its response to each input.

If the input pair yields a frame that matches an existing construction—that is, if the scene-utterance combination corresponds to a reliable association in the model—then the value of $P(k|F)$ will be higher than if no such construction exists. (In the latter case, the best construction is a new one, with low prior probability.) Thus, when comparing the values recorded in response to the appropriate and inappropriate pairing for each utterance, a higher value of $P(k|F)$ corresponds to the child "recognizing" the appropriate scene for the utterance.

Table 2 shows the value of $\log P(k|F)$ across the conditions, after varying amounts of learning (10, 100, and 1000 input pairs; averaged over 10 simulations). The sizable difference between the two (matched and unmatched) pairs for each utterance type mimics the child's ability to pick out the appropriate scene for an utterance based on learned argument structure regularities.

6.3 The Use of Unusual Constructions

Like children, the model mistakenly overgeneralizes, but recovers from these errors only by receiving additional positive evidence (Alishahi and Stevenson, 2005). However, this ability to converge on appropriate argument structures for each verb should not prevent the language learner from making productive generalizations such as *The fly buzzed into the room* (cf. example (2) in Section 1).

Utterance	Scene	# Input Pairs		
		10	100	1000
U_T	S_{CA} (matched)	-5	-5	-5
	S_A (unmatched)	-9	-10	-11
U_I	S_{CA} (unmatched)	-12	-13	-14
	S_A (matched)	-4	-3	-3

Table 2: $\log P(k|F)$ for matched and unmatched scene-utterance pairs.

We test our model with a verb appearing in an unusual (for that verb) construction, to see whether the model can determine appropriate semantic properties. We add a new verb *dance* to the input generation lexicon, with one frame:

head verb		*dance*
verb sem. primitives		⟨*act, manner*⟩
args	roles	⟨*agent*⟩
	categories	⟨*animate*⟩
syntactic pattern		arg1 *verb*

After training on 1000 input pairs, in which *dance* appears only intransitively, we present the model with the following scene-utterance pair:

$\text{DANCE}_{[???]}(\text{KITTY}_{\langle ?\rangle}, \text{DOGGY}_{\langle ?\rangle}, \text{UNDER}_{[]}(\text{TABLE})_{\langle ?\rangle})$
kitty dance doggy under table

The scene representation has been modified to remove the semantic primitives of the verb and the roles of the arguments. Given this partial input, the model must predict multiple missing semantic features, based on the utterance.

Averaged across 10 simulations, the model predicts novel semantic primitives for the verb *dance* in this usage with a probability of .49 for *cause*, .43 for *move*, and negligible probabilities for other primitives. The roles predicted for the arguments, with associated probabilities, are *agent* (.90), *theme* (.94), and *goal* (.99). The model has generalized the feature values of a construction corresponding to the usage of a verb such as *put*, shown above in Figure 1.

We test this ability in sentence production as well, presenting the model with the full scene representation and predicting the most probable syntactic pattern. The sentence *kitty dance doggy under table* is produced as expected.

Unlike overgeneralization errors, the ability of the model to generate novel utterances for unusual situations (or comprehend the meaning of the unusual utterances) does not fade over time by processing more input. Crucially, in these "unusual" cases, the model has not learned a verb-specific frame that sufficiently matches the partial frame to be processed. Therefore, the only reliable knowledge source is a matching construction (if such a construction exists). This property of the model embodies the interaction of entrenchment and statistical preemption in construction use suggested by Goldberg (2005).

7 Related Computational Models

Some recent usage-based models of language acquisition handle syntax/semantics interaction, and generalization to novel situations (Allen, 1997; Niyogi, 2002). For example, Niyogi (2002) proposes a Bayesian model that shows how syntactic and semantic features of verbs interact to support learning. In contrast to our model, the structure of the verb classes and their probabilities, as well as the probabilities of verbs showing particular features, are all fixed. The connectionist model of Allen (1997) is able to make interesting generalizations over argument structure syntax and semantics. However, learning of general constructions is implicit, and the acquired knowledge cannot be used in any language task other than limited comprehension.

Only a few computational models directly address learning of argument structure constructions. Chang (2004) presents a computational model which learns lexical constructions as a mapping between graphical representations of form (typically word order) and meaning (typically role-filler bindings) from annotated child data. Unlike our model, this approach relies on noise-free input and extensive prior knowledge, and constructions are not generalized across verbs. The model of Dominey (2003) learns constructions from narrated video images. The model successfully assigns semantic roles in familiar data, and allows limited generalization of this ability over new verbs. However, in contrast to our approach, learning is highly dependent on the unrealistic assumption of having each form uniquely identify the associated meaning (i.e., forms and meanings are in a one-to-one mapping).

8 Discussion

We have described a Bayesian model for the representation, acquisition and use of argument structure constructions in a usage-based framework. The results of computational experiments with the model demonstrate the feasibility of learning general constructions from individual examples of verb usage, even in the presence of noisy or incomplete input data. We also show that the model can use its acquired construction-based knowledge to generalize to new or low-frequency verbs, or verbs appearing in unusual constructions. These results stem from our novel view of constructions as a mapping of a form to a probability distribution over semantic features, and the corresponding formulation of language learning and use as a Bayesian categorization and prediction process.

Psycholinguistic evidence suggests that children are aware of verb-independent regularities in comprehension much earlier than they can use them in production. Children may begin by learning *weak* constructions which enable only certain kinds of linguistic operations, but become more robust over time (Tomasello and Abbot-Smith, 2002). Our model follows this trend: Constructions emerge early, and can be used in recognizing appropriate scene-utterance inputs as demonstrated here; however, in production, the model generally exhibits an initial strict imitative phase, before a construction is entrenched enough to generalize (Alishahi and Stevenson, 2005).

An important result of our model is that it can account for recovery from overgeneralization, while at the same time allowing for productive generalization of "unusual" constructions. A question that arises is, what are the limitations of using a verb in a novel construction? Although many innovative uses of verbs are acceptable, many others are not. The distinction seems to come from the fundamental semantic properties of the verb; for example, one can say

I hammered the metal flat but not *I played the game finished*. We are currently considering a more sophisticated, fine-grained semantic representation for verbs and scenes, to enable the model to capture these effects.

References

Alishahi, A. and Stevenson, S. (2005). A probabilistic model of early argument structure acquisition. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. to appear.

Allen, J. (1997). Probabilistic constraints in acquisition. In Sorace, A., Heycock, C., and Shillcock, R., editors, *Proc. of GALA97*, pages 300–305.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Bencini, G. M. L. and Goldberg, A. E. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, 43:640–651.

Chang, N. (2004). Putting meaning into grammar learning. In *Proc. of the First Workshop on Psycho-computational Models of Human Language Acquisition*.

Dominey, P. F. (2003). Learning grammatical constructions in a miniature language from narrated video events. In *Proceedings of the 25nd Annual Conference of the Cognitive Science Society*.

Fillmore, C., Kay, P., and O'Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64:501–538.

Fisher, C. (2002). Structural limits on verb mapping: the role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, 5(1):55–64.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:135–176.

Goldberg, A. (2005). *Constructions in Context*. Oxford University Press. to appear.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. The University of Chicago Press.

Lakoff, G. (1987). *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press.

Langacker, R. (1999). *Grammar and conceptualization*. Berlin: Mouton de Gruyter.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, 17:357–374.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. MIT Press.

Rappaport Hovav, M. and Levin, B. (1998). Building verb meanings. In Butt and Geuder, editors, *The Projection of Arguments: Lexical and Computational Factors*, pages 97–134. CSLI Publications.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74:209–253.

Tomasello, M. and Abbot-Smith, K. (2002). A tale of two theories: response to Fisher. *Cognition*, 83:207–214.

Steps Toward Deep Lexical Acquisition

Sourabh Niyogi
Massachusetts Institute of Technology
niyogi@mit.edu

Abstract

I describe steps toward "deep lexical acquisition" based on naive theories, motivated by modern results of developmental psychology. I argue that today's machine learning paradigm is inappropriate to take these steps. Instead we must develop computational accounts of naive theory representations, mechanisms of theory acquisition, and the mapping of naive theories to lexicalizable concepts. This will enable our theories to describe the flexibility of the human conceptual apparatus.

1 Where We Are Now

The present Machine Learning Paradigm

Much of computational linguistics has converged onto a machine learning paradigm that provides us soothing clarity. The machine learning approach defines a problem as a mapping problem – map some acoustic stream onto a list of word tokens, map a list of word tokens onto a parse tree, map a parse tree onto a set of semantic roles or "logical form", map each word in a tree onto its best sense, and so on. We then develop a learning algorithm to accomplish the desired mapping. Multiple groups describe how well their algorithm maps various test sets given various training sets, and describe a "result" to improve upon. The clarity provided by this paradigm is so soothing, one gets the sense we can turn a crank, and indeed, in many cases, progress has been made proceeding precisely along these lines.

Turning the crank on deep lexical acquisition, however, we might feel something is missing. What is it? Underlying any model of deep lexical acquisition is a theory of the *human conceptual apparatus*. Unlike our handle on acoustic streams, word lists, and parse trees, our handle on a suitable "output" for the space of word meanings is remarkably poor. Somehow, via experience (of some kind or another), children acquire a mapping from a space of vocabulary items to a space of lexicalizable concepts – the lexicon; our task as modelers is to figure out how this mapping can occur. Many models for the space of lexicalizable concepts exist: concepts are points in R^n, concepts are Jackendoff's lexical conceptual structures, concepts are FrameNet's frame elements, concepts are Schankian script activators, concepts are distributions over syntactic frames, concepts are grounded in sensorimotor statistics, or all of the above. Almost everyone nowadays reports how their algorithm accomplished some mapping to one or more of these models of concepts. They have to, because today's de facto idea of what constitutes a "result" according the machine learning paradigm today is to do exactly this.

The Golden Oldies formed our concept models

Our models of conceptual spaces did *not* originate from computational linguists following the machine learning paradigm. They were proposed from linguists, psychologists and philosophers back in earlier eras - what we will call Golden Oldies – when the idea of a "result" was somewhat different. There are too many to recall: Quine (1960) argued that the linguist watching the natives uttering *Gavagai!* in the context of a rabbit would nec-

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 91–99,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

essarily require far more constraints than met the eye. Brown (1957) showed that children used syntactic cues to disambiguate between possible meanings; Landau and Gleitman (1985) followed on these insights, showing just how deep it could be, that even blind children could learn *look* and *see*, basing their mapping on syntactic constraints. Chomsky's (1965) notion of "deep structure" – proposed to account for commonplace syntactic phenomena – motivated many insights explored in Gruber (1965)'s thesis, Fillmore (1968)'s classical thematic roles, and Jackendoff (1983)'s Lexical conceptual structures. Hale and Keyser and many linguists labored under the MIT Lexicon project in the 1980s to determine the fundamental features of the lexicon; many of these hard-earned observations appear in Levin (1993). Schank (1972)'s Conceptual dependency theory, Minsky (1975)'s Frames were proposed for the broader goals of capturing commonsense knowledge. Quillian's (1968) and Miller et al (1990)'s WordNet were not intended for models of lexical acquisition or databases to be used in computational linguistics but as models of human semantic memory. Many other Golden Oldies exist, and our debt to them is quite large. Ask what motivates our collection of subcategorization statistics or what drives the quest for semantic roles, and the roots are found in the science questions of the Golden Oldies.

The present Myopic Learning Paradigm

It would have been extremely myopic to take any one of these classical results and accuse their authors of not demonstrating a learning algorithm, not evaluating them on large corpora, and not getting together in workshops to share the results on test sets. The standard for what constituted a result back then consisted of none of these things, because today's machine learning paradigm was just not present then. The questions were:

- Question (1): What is a lexicalizable concept?
- Question (2): How can a word-concept mapping be learned from evidence?

But for reasons that no one really talks about, somehow, the standard of what constitutes a result changed from some balance of Question (1) and (2) to a machine learning paradigm essentially focused on Question (2). The dependency between Question (1) and (2) is quite well-understood, but do we have an adequate answer to (1)? We tell ourselves: *We've gotta build better parsers, speech recognizers, search engines, machine translation systems, so...* let's take shortcuts on Question (1) so as to make progress on Question (2). For many, that shortcut consists of semantic role labels and learning from frame distributions. These shortcuts don't answer Question (1), unfortunately.

2 Where We Need to Go

While the Golden Oldies were used as the foundations of today's lexical acquisition, psychology began to sing a new tune, still balancing Questions (1) and (2).

Children have naive theories

Developmental psychology after the Golden Oldies has shown just how deep our "deep lexical acquisition" theories have to be. On this view, word meanings are couched in changing *naive theories* of how the world works. The model of the child is that the child possesses a naive theory T^* changing state from $T1$ to $T2$, and that there is a space of concepts accessible from $T1$ that substantively different from the space of concepts accessible from $T2$. A learner undergoes *radical conceptual change*. Developmental psychology has not been explicit about the precise form of T^*, nor have they characterized how T^* relates to lexicalizable concepts. But their contributions inform us about the fundamental ingredients of concepts (Question (1)) and inform us what deep lexical acquisition must consist of (Question (2)).

A few examples must suffice in place of a review (c.f. Gopnik and Meltzoff (1997)). Keil (1989)'s transformation studies illustrate theory change in the domain of biology. First, children are shown a picture of a skunk; then, are told a story – that the animal received either (A) surgery or (B) a shot in infancy – and then are shown a picture of a raccoon. Young preschool children judge that the animal is a raccoon, as if they base their judgements on superficial features. Children between 7 and 9 (T2) on the other hand, judge that the raccoon-looking figure in (A) is still a skunk. Adults ($T3$) judge that the raccoon-looking figure in both conditions is still a skunk. Apparently, preschoolers' theory $T1$

lacks the belief that an animal's kind is determined at birth, but this becomes part of the adult's $T3$.

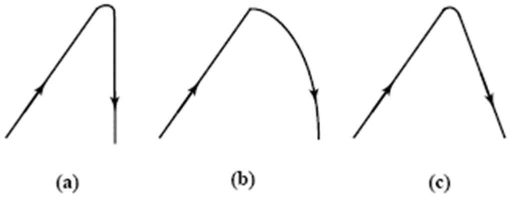
Similarly, preschool children at $T1$ have concept of *death* involving a belief in a continued existence in an alternate location (like sleep); When asked whether dead people dream, eat, defecate, and move, 4 to 6 year olds will say that dead people do all of these, except move (Slaughter et al, 2001). Missing in $T1$ are the causes of death (a total breakdown of bodily functions) and that death is an irreversible, inevitable end. Between 4 and 6, children become superficially aware of the general function of various body parts (e.g "You need a heart to live"). Other phenomena serve the same point: the child at $T1$ thinks *uncle* means friendly middle-aged man, and at $T2$ thinks it means parent's brother. The child at $T1$ thinks *island* means a beachy territory and at $T2$ thinks it means body of land surrounded by water (Keil 1989). And, "theory of mind" concepts/words such as *belief*, *desire*, *wonder*, *pretend* (Wellman and Bartsch 1995, Leslie 2000) are similarly situated.

How "theory-like" $T1$ and $T2$ are is subject to considerable debate (diSessa 1993, Leslie 2000). disessa (1993) describes a large number of causal "p-prims" that are highly context specific and considerably larger in number than what Carey (1985) describes; these are shown to apply to everyday physical phenomena – "force as mover", "vaccuums impel", "overcoming", "springiness", "bigger means lower pitch (or slower)", to name a few. Each of these have a FrameNet-like causal syntax, of some unknown mapping to vocabulary items. Similarly, Rozenblit and Keil (2003) show that non-expert adults have a remarkably superficial notion of how common mechanisms work – such as how a helicopter changes from hovering to forward flight. Theories may be suspiciously weak.

Students have alternative frameworks

Educational psychologists have characterized T^* by asking a different, more practical question: why is it difficult for science students to learn certain scientific concepts (*weight*, *density*, *force*, *heat*, ...) when they come to class? The broad insight is this: students come to class not as blank slates but with alternative *pre*-conceptions that must be understood. Data on their pre-conceptions yields clues as to contents of T^*, well before they walk into science class. Again, a few examples illustrate the point.

Many studies on physics misconceptions have observed deeply held views on the motion of projectiles (McCloskey 1983, Halloun and Hestenes 1985). Ask students to predict what happens when a projectile is thrown upward at an angle, and their answers will typically be consistent with one of (a-c)



These answers are consistent with an "impetus" theory of motion, where an object's motion is exclusively dominated by whatever "impetus" the thrower provides it. Medieval scientists such as Buridan also held similar beliefs; Newtonian mechanics, of course, shows that the answer is a parabola. disessa (1993) report a wider array of these types of physics misconceptions in a theoretical framework.

Likewise, ask students for their knowledge of how their *eyes* work, and they reveal an "extramission" belief: something somehow shoots out from the eye and reaches the objects (Winer et al 2002); they also say that eye is the sole organ in the body responsible for vision. Plato and da Vinci shared these same beliefs. Systematic catalogues of these sorts of observations have been compiled for just about every domain – e.g. megaphones create sounds, heat is a substance, eggs are not alive, the moon and sun are the same size, and so forth (AAAS 1993).

3 What Steps We Must Take

Consider this fascinating phenomena from the Best of Today and the comfort of the grammar-generates-sentence relation will be replaced by queasiness: the terms *theory*, *concept*, and *change* are most unclear, as many developmental psychologists freely admit. But computational linguists may contribute significantly to rendering new clarity: If the Golden Oldies drove the efforts on today's shallow lexical acquisition, the Best of Today's Psychology may drive the results of tomorrow's progress in deep lexical acquisition.

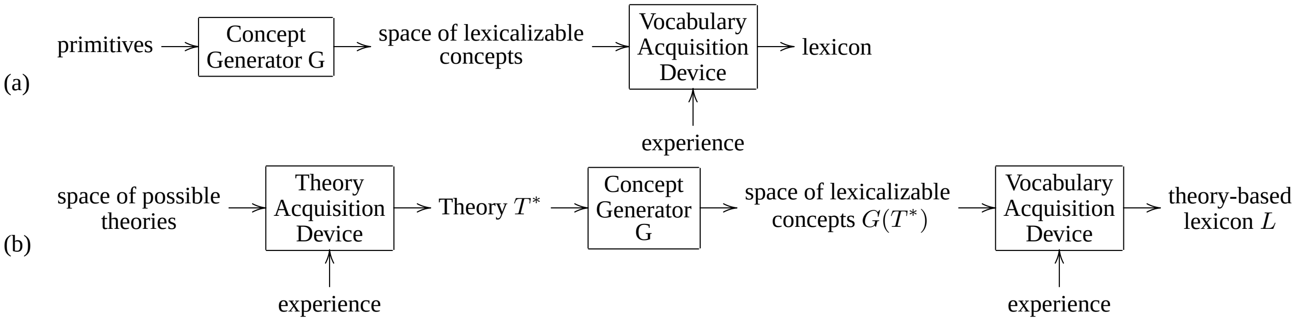

(a)
primitives
Concept Generator G
space of lexicalizable concepts
Vocabulary Acquisition Device
lexicon
experience
(b)
space of possible theories
Theory Acquisition Device
Theory T^*
Concept Generator G
space of lexicalizable concepts $G(T^*)$
Vocabulary Acquisition Device
theory-based lexicon L
experience
experience

Figure 1: (a) The Model of Concepts from the Golden Oldies: used in the present Machine Learning Paradigm; (b) The Universal Theory Model of Concepts: necessary for deep lexical acquisition

The new framework: Universal Theory

We have much progress to make: We can *describe* naive theories precisely; we can *describe* how theory acquisition occurs; we can *describe* the map from naive theories to a set of lexicalizable concepts. We can *describe* how vocabulary acquisition occurs. Figure 1(a) shows the Golden Oldies model of concepts that we must abandon: a Vocabulary Acquisition Device receives a fixed hypothesis space of possible concepts completely determined by a fixed set of primitives; Figure 1(b) shows the *Universal Theory Model of Concepts* that we must take steps towards: A *Theory Acquisition Device* (TAD) outputs a state T^* that describes a learners's naive theory; A *Concept Generator* G maps T^* to a set of lexicalizable concepts $G(T^*)$. A *Vocabulary Acquisition Device* (VAD) uses $G(T^*)$ to learn a lexicon. The theory of the TAD states is *Universal Theory* (UT); a UT metalanguage enables an abstract characterization of *possible theories* – each possible theory describes a system of kinds, attributes, relations, part-whole relations, and causal mechanisms. Within this *Universal Theory Model of Concepts*, we can begin to answer the following core questions:

1. what is the initial state of the TAD?
2. what are possible final states of the TAD?
3. how can the TAD change state?
4. how can the TAD use T^* to parse experience?
5. how does the concept generator G map T^* onto a set of lexicalizable concepts $G(T^*)$?
6. how can the VAD use $G(T^*)$?

We have made progress on these core questions

Many of these questions have been addressed already in computational models where a candidate UT metalanguage and theory T^* is latent. diSessa (1993) catalogs sets of p-prims in naive physics. Atran (1995) describes a theory of family structure. Gopnik et al (2004) uses Bayesian networks to model preschooler's causal reasoning about *blickets*. McClelland and Rogers (2004) describe connectionist models of some of Carey (1985)'s classic results.

In my own work, I have been situating the elements of the Universal Theory Model of Concepts in a microgenesis study, where adult subjects undergo a $T1$ to $T2$ transition (Niyogi 2005). The transition can be understood with a minimal UT metalanguage needed to characterize a set of possible theories: T^* is characterized by a interrelated sets of kinds, attributes, relations, and causal laws. $T1$ and $T2$ are described in that UT metalanguage, and the simplest concept generator G is described that mechanically maps $T1$ and $T2$ onto $G(T1)$ and $G(T2)$. Subjects undergo theory change in a Blocksworld universe (see Figure 2(a)) while learning 3 verbs (*gorp*, *pilk*, *seb*) that refer to the causal mechanisms governing the universe. Subjects interact with a set of 29 blocks, some of which activate other blocks on contact. On activation, subjects are shown a transitive verb frame ("Z is *gorping* L, "U is *sebbing* F", "D is *pilking* Y") in a Word Cue Area. Unbeknownst to subjects, each block belongs to 1 of 4 kinds (A, B, C or D) and 3 activation mechanisms exist between them: lawab: As activate Bs, lawc': Cs activate Cs, and lawd: Ds activate Ds; each of the 3 verbs refers to one the 3 mechanisms. Subjects are probed for the naming conditions on each of the 3 verbs.

Subjects' responses indicate that their TAD state changes from $T^* = T1$ (there is 1 kind of block governed by 1 causal mechanism lawq) to $T^* = T2$

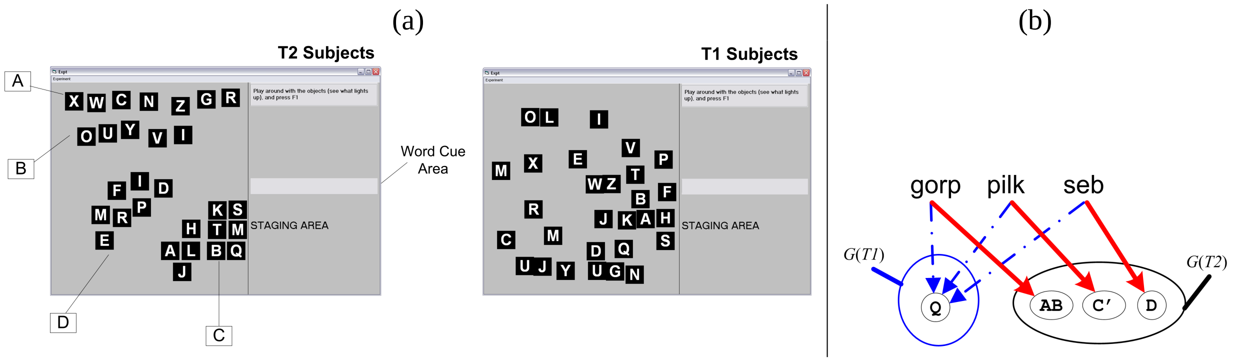


Figure 2: (a) Subjects try to learn the laws and word meanings in a "Causal Blocksworld" computer application by dragging and dropping blocks onto each other. Cues to the meaning of 3 verbs (*gorp*, *pilk* and *seb*) are given in a Word Cue Area. Shown is how two kinds of subjects – $T2$ Subjects and $T1$ Subjects – clustered the blocks; the clusters for the kinds A, B, C and D (boxed) are clear for $T2$ Subjects but no such differentiation is apparent for $T1$ subjects; (b) When $T^* = T1$, all 3 verbs can only be mapped to a single concept in $G(T1) = \{\texttt{Q}\}$ (dashed arrows); When $T^* = T2$, *gorp*, *pilk* and *seb* can be mapped to 3 *new* concepts $\texttt{AB}, \texttt{C}'$ and $\texttt{D}$ in $G(T2)$ (solid arrows).

(there are 4 kinds of blocks governed by 3 distinct causal mechanisms, lawab, lawc' and lawd). But this is not true for all subjects: some remain "T1 subjects" while others move onto become "T2 subjects". Critically, when $T^* = T1$, the verbs can only be mapped to a single concept in $G(T1) = \{\texttt{Q}\}$; When $T^* = T2$, the verbs can be mapped to 3 distinct concepts in $G(T2) = \{\texttt{AB}, \texttt{C}', \texttt{D}\}$ (See Figure 2(b)). Once $T^* = T2$, subjects can "parse" the activation and infer the hidden kind and causal mechanism involved. Critically, subjects cannot learn to distinguish the 3 verbs until $T^* = T2$, when the 3 new concepts emerge in $G(T^*)$. Then *gorp*, *pilk* and *jeb* may be mapped onto those 3 new concepts. These verbs are thus theory-laden in the same way as *death*, *uncle* and *island*.

This UT architecture concretely *dissolves* the Puzzle of Concept Acquisition (Laurence and Margolis 2002): how can a person ever acquire a "new" concept, when a fixed set of primitives exhaustively span the space of possible concepts? Taking the viewpoint of the learner's VAD at a specific moment in time with a *specific* T^*, it has access to *just* those concepts in $G(T^*)$ – acquisition of a new concept is possible if T^* changes. Taking the viewpoint of the learner's *species* across all possible times, the species has access to the *union* of $G(T^*)$ over all possible TAD states – thus a "new" concept for the species is impossible. Which viewpoint one takes is a matter of perspective. Critically, the Golden Oldies model of concepts does not expose the TAD state revealed in the UT model of concepts (Fig. 1a,b).

Universal Theory and the Linguistic Analogy

Computational linguists can progress on these questions, because naive theories are like grammars. Just as a grammar generates a set of possible sentences, a theory T^* generates a set of possible worlds. Just as the space of possible grammars is restricted, so is the space of possible theories. Just as learning a grammar consists of picking a point from a space of possible grammars, learning a theory consists of picking a point from the space of possible theories. The task of writing a naive theory is like writing a grammar. The task of characterizing the space of possible theories requires a theory metalanguage just as characterizing the space of possible grammars requires a grammar metalanguage.

Moreover, research into naive theories does not proceed *separately* from the program of research in grammar. The two programs are bridged by the concept generator G: T^* generates $G(T^*)$, a set of lexicalizable concepts. An adequate account of G would generate concepts present in a particular language, for every language, and for every possible T^*.

Miller et al (1990) distinguish between a constructive and a differential lexicon. In a *differential* theory of the lexicon, meanings can be represented by any symbols that enable a theorist to distinguish among them; In a *constructive theory* of the lexicon, the representation should "contain sufficient information to support an accurate construction of the concept (by either a person or a machine)".

The conceptual analyst who desires to produce a *constructive* theory of the lexicon has four kinds of accounts to provide: (see Niyogi 2005)

- an explanatory account of the space of possible theories, for all persons P
- an explanatory account of the space of possible concepts, for all persons P, for all possible theories
- a descriptive account of a specific theory T^* held by a representative person P (e.g. of a 3-year old or of a 10-year old)
- a descriptive account of a specific lexicon L held by a representative person P (e.g. a 3-year old Chinese speaker, 3-year old English speaker, 10-year old Chinese speaker, 10-year old Chinese speaker)

We may envision a "theory-based lexicon" that would capture the *two* key state variables in Figure 1(b), the two descriptive accounts above: (1) T^* for an idealized human; (2) a set of vocabulary items mapped to points in $G(T^*)$. Very limited instances of a theory-based lexicon can be constructed already for subjects at the end of the experiment – such a theory-based lexicon has (1) $T2$ in the UT metalanguage; (2) the mapping in L to $G(T2)$: *gorp* $=$ AB, *pilk* $=$ C′, *seb* $=$ D. This *constructive* theory-based lexicon would be in stark contrast to *differential* lexicons such as WordNet and FrameNet.

Grounding language in perception is insufficient

Many have proposed deep lexical acquisition by "grounding language in perception" (Siskind 1996, Regier 1996, Roy and Pentland 2002, Yu and Ballard 2004), constructing systems that can learn to utter, e.g. *red, banana*, *hit* and *triangle* in contexts where there are, e.g., three triangles hitting red bananas. Such systems also propose a space of possible concepts exhausted by a *fixed* set of primitives, as in the Golden Oldies model. The initial state of the TAD ($T^*(t = 0)$) can explicitly incorporate all these attributes and relations (contact, luminance, ...); but then, the TAD can *further* change state to yield new kinds, attributes, relations, and causal mechanisms not present in the initial state, but motivated by the data (see Gopnik and Meltzoff 1997). As such, vague appeal to grounding is insufficient; associative processes that may work on *red*, *hit*, *banana*, *eye*, *three* are extremely challenging to generalize to *color, kind, wonder, pilk*, *seb, telescope, maybe* and uninvented *groobles* that cannot be perceived. Again, developmental psychology provides some insight on what theoretical innovations would be required for a suitable interface to sensorimotor apparatus (c.f. Mandler 2004).

Commonsense AI gives UT foundations

Primitives well beyond the sensory apparatus have been developed to describe physical systems qualitatively (Regier 1975, Forbus 1984). They show us some of the possibilities of what T^* and candidate UT metalanguages may look like (quantity spaces, kinds, attributes, relations, part-whole relations, and causal mechanisms that interrelate these sets). Regier (1975)'s description of a *toilet* appears particularly close to Rozenblit and Keil (2003)'s *helicopter*. Later qualitative AI frameworks of Forbus (1984) and Kuipers (1994) may be applied to McCloskey (1982)'s intuitive physics and disessa's (1993) p-prims. Except for the work of Hobbs, Pustejovsky and their colleagues, few have mapped commonsense theories onto the lexicon. Similar domain-general elements of naive math and causality are present in the workds of Hobbs et al (1987), Kennedy and McNally (2002)'s degree representations for gradable predicates, Talmy (1988)'s force dynamics, and the quantity spaces of Kuipers (1994) and Forbus (1984). These disparate frameworks provide foundational elements for a UT metalanguage.

Shortcuts on UT foundations will not work

We must resist the urge to take shortcuts on these foundations. Simply creating slots for foundational phenomena will impede progress. Pustejovsky (1995)'s observations for co-composition have clearly illustrated how much flexibility our interpretation systems must have, e.g. in *He enjoyed the beer/movie*. But specifying the telic role of *beer* and *movie* to be drink and watch does not constitute an adequate theory – we require constraints that relate to the state space of the human conceptual apparatus. Pustejovsky (1995)'s telic, formal, constitutive, agentive roles may be mapped onto T^*'s characterization of artifacts, materials, and so on. We require nothing less than absolute conceptual transparency.

We must bridge UT to analogy

Lakoff and Johnson (1980) and subsequent cognitive linguistics work have catalogued a stunning level of metaphoric usage of language. Lexical extension of items such as *illuminate* in, e.g. *Analogies illuminate us on theory acquisition* are couched in terms of conceptual metaphors such as "ideas are light". Significant steps have been taken to model analogical mapping (c.f. Falkenhainer et al 1989, Bailey et al 1997) and conceptual blending (Fauconnier and Turner 1998). These processes may motivate TAD state changes. In most cases, the the underlying predicates in the source and target domains are ad hocly constructed; a natural source of these predicates may be the sets internal to T^* (kinds, attributes, relations, causal mechanisms); similarity between domains may be determined by the structural properties of the UT metalanguage and G. If T^* incorporates the common causal mechanisms behind ideas and light transmission, for example, then one may strive for a shorter lexicon where the vocabulary item *illuminate* happens to be used in both domains with "one" core entry. An adequate theory of this process would obviously reduce the number of so called "senses" in word sense disambiguation.

4 What We Assumed Wrong

Modern computational linguistics appears to have made a set of assumptions that deserve reanalysis, given the availability of other options.

Assumption: A fixed alphabet of meaning components exists, and we know what it is

A key assumption dating to the Golden Oldies is that the meaning of a sentence is adequately captured by a "logical form" (LF) characterized by a *fixed alphabet* of meaning components (e.g. thematic roles, lexical semantic primitives, conceptual dependency primitives). Today's computational linguistics program uses this assumption to demonstrate systems that answer "who did what to whom, where, why, . . ." questions, given sentences like:

John saw the man with the telescope.

John hit the man with the umbrella.

Is the computational linguist is expected to be satisfied when systems can answer *Who saw the man with the telescope?* or *Who did John hit with the umbrella?* This year's CoNLL Shared Task, mapping sentences onto semantic roles, assumes the above. But try these: Does John have eyes? Were they ever open when he was looking through the telescope? Could John know whether the man was wearing underwear? Did the umbrella move? Did John move? Did the man feel anything when he was hit? Was John alive? Was the man alive? Why would John need a telescope to see the man, when he has eyes? Why would John use an umbrella when his hands would do? Something is missing in these systems.

We should be more accountable. Developmental psychology showed that theory change and conceptual change is possible, proving this assumption is *wrong*: the alphabet behind sentence meaning is a *varying* set of lexicalizable concepts $G(T^*)$. Missing in today's systems attaching AGENT (or FrameNet's Perceiver_passive, or Impactor) to *John* and INSTRUMENT to *umbrella* and *telescope* is T^*, and a mapping of the lexical items to $G(T^*)$. What T^* must contain, in some as yet unknown form, is a T of physics described by McCloskey and disessa (1993), a T of vision studied by Landau and Gleitman (1985) and Winer et al (2002), a T of body studied by Carey (1985), a T of materials and artifacts studied by Hobbs et al (1987) and Pustejovsky (1995). This T^*, when mapped via G, forms the alphabet of the above 2 sentences.

Assumption: The machine learning paradigm can treat deep lexical acquisition.

If we reject the assumption that there is some "meaning" of a sentence spanned by a set of meaning primitives, the soothing clarity of the machine learning paradigm is no longer available. We cannot map parse trees onto sentence meanings. The possibility of "Putting Meaning in Your Trees" (Palmer 2004) completely disappears. We may still use the machine learning paradigm to parse, disambiguate and recognize speech. But these results are of little use to model theory, concept and lexical acquisition, because there is no output representation where a suitable training set could be collected. The human conceptual apparatus is not that simple: the VAD requires $G(T^*)$ (which changes, as T^* changes), and for *that* we need explanatory accounts of UT and G, and must recognize the diverse ways the TAD may change state.

Assumption: Paths from shallow to deep lexical acquisition exist

The Golden Oldies Models of concepts (Figure 1a) and the Universal Theory models of concepts (Figure 1b) are *incommensurable*. The path from the shallow to the deep cannot be declared to exist by fiat. Wishful thinking is inappropriate, because one architecture is more powerful than the other: the Golden Oldies model did not expose the TAD state space. Instead, lexical semantics results obtained under the Golden Oldies model require translation into the UT model: the privileged position syntactic positions that motivated thematic roles and lexical semantics primitives, the bi-partite event structure revealed through adverbial modification, and so on. This translation is mediated in G, and will not yield a notational variant of what we started with.

Assumption: Verb classes determine meanings

We must distinguish between a representation of verb meanings *determined by* the distribution of subcategorization frames and *cued by* these frames. Landau and Gleitman (1990) showed that verb's participation in some frames but not others are *cues* that a child uses to constrain verb meaning. Levin and Rappaport-Hovav (1998) explicitly distinguish *structural* and *idiosyncratic* components of meaning. But neither claim that verb classes or statistical distributions of subcategorization frames *determine* verb meaning. Yet VerbNet maps verbs to predicates in precisely this way: (Kingsbury et al 2002).

cure, *rob*, . . .: Verbs of Inalienable Possession cause(Agent,E) location(start(E),Theme,Source)
marry, *divorce*, . . .: Verbs of Social Interaction social_interaction(. . .)

The distinction between *cure* and *rob*, or between *marry* and *divorce* is not astonishing to the English speaker. Causal mechanisms behind disease, possession, and the marital practices that were labeled *idiosyncratic* by the lexical semanticist must be captured in T^*.

Assumption: Language is separate from general systems of knowledge and belief

This "defining" assumption helped for the Golden Oldies, but innovations in developmental psychology motivate dropping this assumption. The bridge is provided by the concept generator G: it maps a naive theory T^* (general systems of knowledge and belief) to $G(T^*)$, used by the VAD (language).

Assumption: Real-world knowledge is Bad

The absence of the soothing clarity of the machine learning paradigm and presence of real world knowledge in T^* brings forth 2 associations:

Early Schank/Cyc = Much Knowledge = UT research = Bad
Statistics = Little Knowledge = shallow semantics = Good

The associations lead to the inference that Universal Theory research will suffer a similar fate as the 70s Schankian program and the Cyc program (Schank 1972, Lenat and Guha 1990). However, this inference is incorrect. The 70s Schankian program and Cyc efforts did not carefully consider the constraints of syntactic phenomena or developmental psychology. Schank and his colleagues stimulated research in qualitative physics and explanation-based learning that addressed many of these deficiencies, but there is much work to be done to bridge today's efforts in deep lexical acquisition to this.

Assumption: Others will provide us the answers

Lexical semanticists now rely on cognitive explanations far more heavily than ever before. Jackendoff (2002) concludes: "someone has to study all these subtle frameworks of meaning - so why not linguists?" Levin and Rappaport-Hovav (2003), addressing denominal verbs such as *mop* and *butter*, now freely point to "general cognitive principles" rather than situate knowledge in the lexicon. Rather than consume lexical semantics of the Golden Oldies, we can draw upon our toolbox to again answer Question (1): "what is a lexicalizable concept?"

5 We Must Change Our Concepts

Stop working with models of concepts from the Golden Oldies. Start questioning whether results under the machine learning paradigm are really *results*. Change your concept of a *result*. Learn how children do theory, concept and vocabulary acquisition. Expose the fundamental ingredients of concepts. Change your concept of *deep*. Change your concept of *computational linguistics*. Radical conceptual change is possible. Write some new songs, and sing some new tunes. We can have some Great Golden Oldies of Tomorrow.

References

S. Atran. Classifying nature across cultures. In E. Smith and D. Osherson, editors, *Thinking: An invitation to cognitive science*, Cambridge, MA, 1995. MIT Press.

D. Bailey, J. Feldman, S. Narayanan, and G. Lakoff. Modeling embodied lexical development. In *Proceedings of the Annual Cognitive Science Society*, 1997.

K. Bartsch and H. Wellman. *Children Talk about the Mind*. Oxford University Press, New York, 1995.

R. Brown. Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology*, 1957.

S. Carey. *Conceptual Change in Childhood*. MIT Press, Cambridge, MA, 1985.

N. Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

A. M. Collins and M. R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–247, 1969.

B. Falkenhainer, K. Forbus, and D. Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41:1–63, 1989.

G. Fauconnier and M. Turner. Conceptual integration networks. *Cognitive Science*, 22(2):133–187, 1998.

C. Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90, New York, 1968. Holt, Rinehart and Winston.

C. Fillmore, C. Wooters, and C. Baker. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*, Hong Kong, 2001.

K. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

D. Gentner. Why we're so smart. In D. Gentner and S. Goldin-Meadow, editors, *Language in mind: Advances in the study of language and thought*, pages 195–235, Cambridge, MA, 2003. MIT Press.

A. Gopnik, C. Glymour, D. Sobel, L. Schultz, and T. Kushnir. Theory formation and causal learning in children: Causal maps and bayes nets. *Psychological Review*, in press.

A. Gopnik and A. Meltzoff. *Words, thoughts and theories*. MIT Press, Cambridge, MA, 1997.

J. Hobbs, W. Croft, T. Davies, D. Edwards, and K. Law. Commonsense metaphysics and lexical semantics. *Computational Linguistics*, 13:241–250, 1987.

R. S. Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, MA, 1983.

R. S. Jackendoff. *Foundations of Language*. Oxford University Press, Oxford, 2002.

F. Keil. *Semantic and Conceptual Development: An Ontological Perspective*. Harvard University Press, Cambridge, MA, 1979.

C. Kennedy and L. McNally. Scale structure and the semantic typology of gradable predicates. *Language*, 2002.

P. Kingsbury, M. Palmer, and M. Marcus. Adding semantic annotation to the penn treebank. In *Proceedings of Human Language Technology Conference*, 2002.

B. Kuipers. *Qualitative Reasoning*. MIT Press., Cambridge, MA, 1994.

B. Landau and L. R. Gleitman. *Language and experience: Evidence from the blind child*. Harvard University Press, Cambridge, MA, 1985.

S. Laurence and E. Margolis. Radical concept nativism. *Cognition*, 86:22–55, 2002.

D. Lenat and D. Guha. *Building large knowledge-based systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Reading, MA, 1990.

A. Leslie. How to acquire a representational theory of mind. In D. Sperber, editor, *Metarepresentations: An Multidisciplinary perspective.*, pages 197–223, Oxford, 2000. Oxford Press.

B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, 1993.

B. Levin and M. Rappaport-Hovav. Objecthood and object alternations. *ms*, 2003.

J. Mandler. *Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press, New York, 2004.

M. McCloskey. Intuitive physics. *Scientific American*, 248:122–130, 1983.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on wordnet. *International Journal of Lexicology*, 3(4), 1990.

M. Minsky. A framework for representing knowledge. In P. Winston, editor, *The psychology of Computer Vision.*, pages 211–277, New York, 1975. McGraw-Hill.

N. Nersessian. Comparing historical and intuitive explanations of motion: Does naive physics have a structure? In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, pages 412–420, 1989.

S. Niyogi. Aspects of the logical structure of conceptual analysis. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 2005.

S. Niyogi. The universal theory model of concepts and the dissolution of the puzzle of concept acquisition. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 2005.

M. Palmer. Putting meaning in your trees. In *CoNLL-2004*, 2004.

J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

W. Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

T. Regier. *The Human Semantic Potential*. MIT Press, Cambridge, MA, 1996.

T. Rogers and J. McClelland. *Semantic Cognition: A parallel distributed Processing approach*. MIT Press, Cambridge, MA, 2004.

D. Roy and Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26:113–146, 2002.

L. Rozenblit and F. Keil. The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26:521–562, 2002.

R. Schank. Conceptual dependency theory. *Cognitive Psychology*, 3:552–631, 1972.

J. Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91, 1996.

V. Slaughter, R. Jaakola, and S. Carey. Constructing a coherent theory: children's biological understanding of life and death. In M. Siegel and C. Peterson, editors, *Children's understanding of biology and health*, Cambridge, 1999. Cambridge University.

V. Slaughter, R. Jaakola, and S. Carey. Constructing a coherent theory: children's biological understanding of life and death. In M. Siegel and C. Peterson, editors, *Children's understanding of biology and health*, Cambridge, 1999. Cambridge University.

C. Yu and Dana H. Ballard (2004) A Unified Model of Early Word Learning: Integrating Statistical and Social Cues. *Proceedings of the 3rd International Conference on Development and Learning*, 2004.

Author Index