

EACL 2009

**Proceedings
of the
EACL 2009 Workshop on
Cognitive Aspects
of
Computational
Language Acquisition**

31 March 2009

Megaron Athens International Conference Centre

Athens, Greece

Production and Manufacturing by
TEHNOGRAFIA DIGITAL PRESS
7 Ektoros Street
152 35 Vrilissia
Athens, Greece

©2009 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Preface

This volume contains the papers accepted for presentation at the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition, held in Athens, Greece on the 31st of March, 2009. This workshop is the second of a series which was initiated during ACL 2007, held in Prague. The first edition of the workshop was organised by Anna Korhonen, Paula Buttery and Aline Villavicencio.

The past decades have seen a massive expansion in the application of statistical and machine learning methods to natural language processing (NLP). This work has yielded impressive results in numerous speech and language processing tasks including speech recognition, morphological analysis, parsing, lexical acquisition, semantic interpretation, and dialogue management. Advances in these areas are generally viewed as engineering achievements, but recently researchers have begun to investigate the relevance of computational learning techniques to research on human language acquisition. These investigations have double significance since an improved understanding of human language acquisition will not only benefit cognitive sciences in general, but may also feed back to the NLP community, placing researchers in a better position to develop new language models and/or techniques.

Success in this type of research requires close collaboration between NLP and cognitive scientists. The aim of this workshop is thus to bring together researchers from the diverse fields of NLP, machine learning, artificial intelligence, linguistics, psycho-linguistics, etc. who are interested in the relevance of computational techniques for understanding human language learning. The workshop is intended to bridge the gap between the computational and cognitive communities, promote knowledge and resource sharing, and help initiate interdisciplinary research projects.

In the call for papers we solicited papers describing cognitive aspects of computational language acquisition. The programme committee has selected 7 papers for publication that are representative of the state-of-the-art in this interdisciplinary area. Each full-length submission was independently reviewed by three members of the program committee, who then collectively faced the difficult task of selecting a subset of papers for publication from a very strong field.

We would like to thank our two invited speakers, Massimo Poesio and Robert Berwick, all the authors who submitted papers, as well as the members of the programme committee for the time and effort they contributed in reviewing the papers. Our thanks go also to the organisers of the main conference, the publication chairs, and the conference workshop committee headed by Miriam Butt and Stephen Clark.

Afra Alishahi, Thierry Poibeau and Aline Villavicencio

Organizers

Organizers:

Afra Alishahi, University of Saarland (Germany)
Thierry Poibeau, CNRS and University Paris 13 (France)
Aline Villavicencio, Federal University of Rio Grande do Sul (Brazil) and University of Bath (UK)

Program Committee:

Colin J Bannard, Max Planck Institute for Evolutionary Anthropology (Germany)
Marco Baroni, University of Trento (Italy)
Robert C. Berwick, Massachusetts Institute of Technology (USA)
Jim Blevins, University of Cambridge (UK)
Rens Bod, University of Amsterdam (Netherlands)
Antal van den Bosch, Tilburg University (Netherlands)
Chris Brew, Ohio State University (USA)
Ted Briscoe, University of Cambridge (UK)
Robin Clark, University of Pennsylvania (USA)
Stephen Clark, University of Oxford (UK)
Matthew W. Crocker, Saarland University (Germany)
James Cussens, University of York (UK)
Walter Daelemans, University of Antwerp (Belgium) and Tilburg University (Netherlands)
Ted Gibson, Massachusetts Institute of Technology (USA)
Henriette Hendriks, University of Cambridge (UK)
Julia Hockenmaier, University of Illinois at Urbana-Champaign (USA)
Marco Idiart, Federal University of Rio Grande do Sul (Brazil)
Mark Johnson, Brown University (USA)
Aravind Joshi, University of Pennsylvania (USA)
Anna Korhonen, University of Cambridge (UK)
Alessandro Lenci, University of Pisa (Italy)
Massimo Poesio, University of Trento (Italy)
Brechtje Post, University of Cambridge (UK)
Ari Rappoport, The Hebrew University of Jerusalem (Israel)
Dan Roth, University of Illinois at Urbana-Champaign (USA)
Kenji Sagae, University of Southern California (USA)
Sabine Schulte im Walde, University of Stuttgart (Germany)
Mark Steedman, University of Edinburgh (UK)
Suzanne Stevenson, University of Toronto (Canada)
Patrick Sturt, University of Edinburgh (UK)
Bert Vaux, University of Wisconsin (USA)
Charles Yang, University of Pennsylvania (USA)
Menno van Zaanen, Macquarie University (Australia)
Michael Zock, LIF, CNRS, Marseille (France)

Invited Speakers:

Robert Berwick, Massachusetts Institute of Technology (USA)
Massimo Poesio, University of Essex (UK) and University of Trento (Italy)

Table of Contents

<i>Towards a Formal View of Corrective Feedback</i> Staffan Larsson and Robin Cooper	1
<i>A Collaborative Tool for the Computational Modelling of Child Language Acquisition</i> Kris Jack	10
<i>What's in a Message?</i> Stergos Afantenos and Nicolas Hernandez	18
<i>Another Look at Indirect Negative Evidence</i> Alexander Clark and Shalom Lappin	26
<i>Categorizing Local Contexts as a Step in Grammatical Category Induction</i> Markus Dickinson and Charles Jochim	34
<i>Darwinised Data-Oriented Parsing - Statistical NLP with Added Sex and Death</i> Dave Cochran	42
<i>Language Diversity across the Consonant Inventories: A Study in the Framework of Complex Networks</i> Monojit Choudhury, Animesh Mukherjee, Anupam Basu, Niloy Ganguly, Ashish Garg and Vaibhav Jalan	51

Conference Program

Tuesday, March 31, 2009

Session 1: Introduction

- 9:20–9:30 Opening Remarks
- 9:30–10:30 Invited Talk by Massimo Poesio
- 10:30–11:00 Coffee Break

Session 2: Theoretical and Practical Aspects of Language Acquisition

- 11:00–11:30 *Towards a Formal View of Corrective Feedback*
Staffan Larsson and Robin Cooper
- 11:30–12:00 *A Collaborative Tool for the Computational Modelling of Child Language Acquisition*
Kris Jack
- 12:00–12:30 *What's in a Message?*
Stergos Afantenos and Nicolas Hernandez
- 12:30–14:00 Lunch Break

Session 3: Learnability and Grammatical Inference

- 14:00–15:00 Invited Talk by Robert Berwick
- 15:00–15:30 *Another Look at Indirect Negative Evidence*
Alexander Clark and Shalom Lappin
- 15:30–16:00 *Categorizing Local Contexts as a Step in Grammatical Category Induction*
Markus Dickinson and Charles Jochim
- 16:00–16:30 Coffee Break

Tuesday, March 31, 2009 (continued)

Session 4: The Ecology of Language

16:30–17:00 *Darwinised Data-Oriented Parsing - Statistical NLP with Added Sex and Death*
Dave Cochran

17:00–17:30 *Language Diversity across the Consonant Inventories: A Study in the Framework of Complex Networks*
Monojit Choudhury, Animesh Mukherjee, Anupam Basu, Niloy Ganguly, Ashish Garg
and Vaibhav Jalan

17:30–17:45 Closing Remarks

Invited Talks

Conceptual Descriptions: Evidence from Corpora, the Mind, and the Brain

Massimo Poesio

All too often work in computational linguistics on the acquisition of conceptual descriptions takes place in isolation from work on concepts in psychology and neural science. We feel this is a mistake as evidence from these related disciplines can provide us with better ways of evaluating our results. In the talk I will present work in CIMEC on using cognitive evidence to evaluate the results of lexical acquisition work - specifically, using feature norms to evaluate the acquisition of features, and using EEG data to evaluate the results of categorization experiments.

Joint work with Marco Baroni, Brian Murphy, Eduard Barbu, and Abdulrahman Almuhareb, among others.

Trebank Parsing and Knowledge of Language: A Cognitive Perspective

Robert C. Berwick

Over the past 15 years, there has been increasing use of linguistically annotated sentence collections such as the LDC Penn Tree Bank (PTB) for constructing statistically based parsers. While these parsers have generally been built for engineering purposes, more recently such approaches have been advanced as potential cognitive solutions, e.g., for the problem of human language acquisition. Here we examine this possibility critically: we assess how well these Treebank parsers actually approach human/child language competence. We find that such systems fail to replicate many, perhaps most, empirically attested grammaticality judgments; seem overly sensitive, rather than robust, to training data idiosyncrasies; and easily acquire unnatural syntactic constructions never attested in human languages. Overall, we conclude that existing statistically based treebank parsers fail to incorporate much knowledge of language in these three senses. We discuss the implications of these results for the improvement of Treebank parsers and their cognitive relevance.

Joint work with Professor Sandiway Fong, University of Arizona.

Towards a formal view of corrective feedback

Staffan Larsson and Robin Cooper

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{sl, cooper}@ling.gu.se

Abstract

This paper introduces a formal view of the semantics and pragmatics of corrective feedback in dialogues between adults and children. The goal of this research is to give a formal account of language coordination in dialogue, and semantic coordination in particular. Accounting for semantic coordination requires (1) a semantics, i.e. an architecture allowing for dynamic meanings and meaning updates as results of dialogue moves, and (2) a pragmatics, describing the dialogue moves involved in semantic coordination. We illustrate the general approach by applying it to some examples from the literature on corrective feedback, and provide a fairly detailed discussion of one example using TTR (Type Theory with Records) to formalize concepts. TTR provides an analysis of linguistic content which is structured in order to allow modification and similarity metrics, and a framework for describing dialogue moves and resulting updates to linguistic resources.

1 Introduction

Here are a few examples of corrective feedback:

A: That's a nice bear.

B: Yes, it's a nice panda.

Abe: I'm trying to tip this over, can you tip it over? Can you tip it over?

Mother: Okay I'll turn it over for you.

Adam: Mommy, where my plate?

Mother: You mean your saucer?

Naomi: Birdie birdie.

Mother: Not a birdie, a seal.

Naomi: mittens.

Father: gloves.

The first one is made up, the others are quoted from various sources in (Clark and Wong, 2002) and (Clark, 2007). In general, corrective feedback can be regarded as offering an alternative form to the one that the speaker used. We are interested in interactions such as these since we believe that dialogue interaction plays an important role in establishing a shared language, not only in first (or second) language acquisition but also in the coordination of meaning in adult language, in historical language change, and in language evolution.

Two agents do not need to share exactly the same linguistic resources (grammar, lexicon etc.) in order to be able to communicate, and an agent's linguistic resources can change during the course of a dialogue when she is confronted with a (for her) innovative use. For example, research on alignment shows that agents negotiate domain-specific microlanguages for the purposes of discussing the particular domain at hand (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Pickering and Garrod, 2004; Brennan and Clark, 1996; Healey, 1997; Larsson, 2007). We will use the term *semantic coordination* to refer to the process of interactively coordinating the meanings of linguistic expressions.

This paper presents work towards a formal theory of corrective feedback, and semantic coordination in general. It takes a view of natural languages as toolboxes for constructing domain-specific microlanguages, and provides an analysis of linguistic content which is structured in order to allow modification of, and similarity metrics over, meanings.

According to (Cooper and Ranta, 2008), a "language" such as Swedish or English is to be

regarded as a collection of resources (a “tool-box”) which can be used to construct local micro-languages. We take the view that speakers of natural languages are constantly in the process of creating new language to meet the needs of novel situations in which they find themselves.

Accounting for corrective feedback requires (1) dynamic representations of concepts which can be modified in various ways, in a process of semantic coordination, and (2) a description of dialogue strategies involved in semantic coordination.

Accordingly, the research effort which the work presented here is part of aims towards an account of semantic coordination in dialogue, consisting of two parts:

1. semantics: an account of how meanings (and concepts) can be updated
2. pragmatics: an account of how meanings (and concepts) are coordinated in dialogue and how dialogue moves governing coordination are related to semantic updates

These parts will be presented below, starting with the pragmatics. At the end of the paper, we will step back and consider the implications of our tentative results.

2 The pragmatics of corrective feedback

To get a handle on the pragmatic processes involved in corrective feedback, we will show how such interactions can be analysed in terms of dialogue moves related to semantic updates. This approach builds on, and extends, the Information State Update approach to dialogue management (Traum and Larsson, 2003).

2.1 A taxonomy of corrective feedback

Below, we classify our examples into four kinds of corrective feedback.

- Example 1: “In-repair”
 - Abe: I’m trying to tip this over, can you tip it over? Can you tip it over?
 - Mother: Okay I’ll turn it over for you.
- Example 2: Clarification request
 - Adam: Mommy, where my plate?
 - Mother: You mean your saucer?
- Example 3: “Explicit replace”

- Naomi: Birdie birdie.
- Mother: Not a birdie, a seal.

- Example 4: “Bare” correction

- Naomi: mittens.
- Father: gloves.

2.2 Dialogue moves for corrective feedback

We will now introduce a representation of dialogue moves used in corrective feedback. The general format we will use is

- **offer-form:** *TYPE*(*ARGS*)

where *ARGS* may include one or several of the following:

- proposed form (*P* below)
- replaced form (*R* below)
- sentence frame (*F* below)

In the representation above, *TYPE* is one of the following, corresponding to the kinds of corrective feedback distinguished above:

- **in-repair**
- **cr**
- **explicit-replace**
- **bare**

In-repair The in-repair type of corrective feedback takes two arguments, the proposed form and a sentence frame. It is generally preceded by an utterance containing the sentence frame applied to the replaced form.

- **offer-form:in-repair**(*P*, *F*)

For illustration, let’s look again at our example, now with typography indicating PROPOSED FORM, **replaced form** and *sentence frame*:

A(be): Can you **tip** *it over*?
 M(other): Okay I’ll **TURN** *it over* for you.

In relation to *A*’s utterance, *M*’s utterance contains the same sentence frame *F*, roughly “[*Mother*] *_ it over*”. However, they differ in that whereas *M*’s utterance has the proposed word *P* = “TURN”, *A*’s utterance has *R* = “**tip**”. If we

say that sentence frames can be applied to expressions, resulting in the “_” in the frame being replaced with the expression (much as in lambda reduction), we can say that *A*’s utterance has the form “ $F(R)$ ” = “[*Mother*] _ *it over*”(“**turn**”) = “[*Mother*] **turn** *it over*” whereas *M*’s utterance has the form “ $F(R)$ ”. *M*’s utterance corresponds to the dialogue move:

offer-form:in-repair(“turn”, “[*M*] _ [*it*] over”)

Note that the syntactic parallelism is not complete; we have ignored the complication that one utterance contains “can” and the other “will” (in reduced form). The notion of sentence frame used here is a simplification of a more complex relation of syntactic and semantic parallelism which needs to be further explored.

Note also that in addition to providing corrective feedback, *M*’s utterance also accepts the content of the previous contribution. Note that *M* might instead have said something like “No, but I’ll *turn* it over for you”.

Clarification requests As in the case of in-repair offers, offers involving clarification requests also provide the proposed form together with a sentence frame linking the move to a previous utterance by the child; presuming that the latter has the form “... $F(R)$ ”, the offer can be represented as

offer-form:cr(P, F)

Let’s revisit our example, making explicit the P , F and R parameters:

A(dam): Mommy, where *my* **plate**?

M(other): You mean *your* SAUCER?

Here, we have F = “[*Adam’s*] _”, R = “**plate**” and P = “SAUCER”. Accordingly, we can describe *M*’s utterance as a dialogue move:

offer-form:cr(“saucer”, “[*A’s*] _”)

Typically, CRs have the interpretation “you mean/want $F(P)$?”. In addition to offering an alternative form P of expression, a clarification request also explicitly raises the issue whether the offer of P is accepted, and is typically followed by a positive (or negative) answer by the child.

Note that CRs, as well as some other types of offers, may not be intended as corrections by

the adult, but simply as attempts at understanding what the child wants to communicate. The crucial point for our purposes here is the effect these moves have on the addressee, rather than the underlying intention. In general, if I learn something from someone else, it may not be of great importance for my learning if they intended for me to learn it or not.

Explicit replace In contrast to in-repairs and clarification requests, explicit offers of replacements need not rely on sentence frames to figure out the replaced form, as it is (as the name indicates) explicitly mentioned in the offer.

N(aomi): Birdie **birdie**

M(other): Not a **birdie**, a SEAL

We represent this kind of dialogue move thus:

offer-form:explicit-replace(P, R)

In the example, the move is **offer-form:explicit-replace**(“seal”, “birdie”). Explicit replace offers are preceded by an utterance consisting of or containing the replaced form R , and typically have the form “(that’s) not DET R , (that’s) DET P ” or similar.

Explicit replace offers differ from in-repairs and clarification requests by clearly signalling that the replaced form is not appropriate, and by being clearly intended as providing corrections rather than (just) figuring out what the child is trying to communicate.

Bare offers Bare offers are the formally simplest kind of corrective feedback, consisting simply of the proposed form.

Naomi: **Mittens**

Father: GLOVES.

The dialogue move representation is

offer-form:bare(P)

In the example, the move is **offer-form:bare**(“gloves”). Since neither sentence frame or replaced form is provided, the replaced form must be figured out from the conversational situation as a whole. Just as explicit replace offers, bare offers are primarily intended as providing

corrections.

2.3 Generalising the dialogue move representation

The different variants for corrective feedback all do basically the same work; they indicate that the child needs to modify his or her take on the meaning of the proposed term, and perhaps also the replaced term. A possible difference is that some forms more clearly provide evidence that the replaced form is not appropriate, whereas others leave this open. Ignoring this complication for the moment, we can provide a general form for the various types of offers of new forms, with the proposed form and the replaced form as arguments:

offer-form(P, R)

Using this representation, the dialogue move analyses above can be reformulated as, in the order they appear above:

- **offer-form**(“turn”, “tip”)
- **offer-form**(“saucer”, “plate”)
- **offer-form**(“seal”, “birdie”)
- **offer-form**(“gloves”, “mittens”)

In moves which do not explicitly indicate the replaced form R , contextual interpretation involves chart alignment and reasoning about active edges (represented here by the sentence frame) to locate an expression R parallel to P in the previous utterance.

2.4 Agents that coordinate resources

As in the information state update approach in general, dialogue moves are associated with information state updates. For semantic coordination, the kind of update is rather different from the one associated with dialogue moves for coordinating on task-related information, and involves updating the linguistic resources available to the agent (grammar, lexicon, semantic interpretation rules etc.), rather than e.g. the conversational scoreboard as such. Our view is that agents do not just have monolithic linguistic resources as is standardly assumed. Rather they have generic resources which they modify to construct local resources for sublanguages for use in specific situations. Thus an agent A may associate a linguistic expression c with a particular concept (or collection of concepts if c is ambiguous) $[c]^A$ in

its generic resource. In a particular domain α c may be associated with a modified version of $[c]^A$, $[c]_\alpha^A$. In some cases $[c]_\alpha^A$ may contain a smaller number of concepts than $[c]^A$, representing a decrease in ambiguity. Particular concepts in $[c]_\alpha^A$ may be a *refinement* of one in $[c]^A$, that is, the domain related concepts have an extension which is a proper subset of the extension of the corresponding generic concept. This will, however, not be the case in general. For example, a black hole in the physics domain is not normally regarded as an object described by the generic or standard meaning of *black hole* provided by our linguistic resources outside the physical domain. Similarly a variable in the domain of logic is a syntactic expression whereas a variable in experimental psychology is not and quite possibly the word *variable* is not even a noun in generic linguistic resources.

Our idea is that the motor for generating new such local resources in an agent lies in coordinating resources with another agent in a particular communicative situation s . The event s might be a turn in a dialogue, as in the examples we are discussing in this paper, or, might, for example, be a reading event. In a communicative situation s , an agent A may be confronted with an *innovative* utterance c , that is, an utterance which either uses linguistic expressions not already present in A 's resources or linguistic expressions known by A but associated with an interpretation distinct from that provided by A 's resources. At this point, A has to accommodate an interpretation for c which is specific to s , $[c]_s^A$, and which may be anchored to the specific objects under discussion in s .

Whereas in a view of semantics inherited from formal logic there is a pairing between a linguistic expression c and an interpretation c' (or a set of several interpretations if c is ambiguous), we want to see c as related to several interpretations: $[c]_s^A$ for communicative situations s , $[c]_\alpha^A$ for domains α (where we imagine that the domains are collected into a complex hierarchy or more and less general domains) and ultimately a general linguistic resource which is domain independent, $[c]^A$. We think of the acquisition of a pairing of an expression c with an interpretation c' as a progression from an instance where c' is $[c]_s^A$ for some particular communicative situation s , through potentially a series of increasingly general domains α where c' is regarded as being one of the interpretations in $[c]_\alpha^A$ and finally arriving at a state where

c' is associated with c as part of a domain independent generic resource, that is, c' is in $[c]^A$. There is no guarantee that any expression-interpretation pair will survive even beyond the particular communicative situation in which A first encountered it. For example, the kind of *ad hoc* coinages described in (Garrod and Anderson, 1987) using words like *leg* to describe part of an oddly shaped maze in the maze game probably do not survive beyond the particular dialogue in which they occur. The factors involved in determining how a particular expression-interpretation pair progresses we see as inherently stochastic with parameters including the degree to which A regards their interlocutor as an expert, how many times the pairing has been observed in other communicative situations and with different interlocutors, the utility of the interpretation in different communicative situations, and positive or negative feedback obtained when using the pairing in a communicative situation. For example, an agent may only allow a pairing to progress when it has been observed in at least n different communicative situations at least m of which were with an interlocutor considered to be an expert, and so on. We do not yet have a precise proposal for a theory of these stochastic aspects but rather are seeking to lay the groundwork of a semantic treatment on which such a theory could be built.

3 The semantics of corrective feedback

3.1 Representing concepts using TTR

We shall make use of type theory with records (TTR) as characterized in Cooper (2005; 2008) and elsewhere. The advantage of TTR is that it integrates logical techniques such as binding and the lambda-calculus into feature-structure like objects called record types. Thus we get more structure than in a traditional formal semantics and more logic than is available in traditional unification-based systems. The feature structure like properties are important for developing similarity metrics on meanings and for the straightforward definition of meanings modifications involving refinement and generalization. The logical aspects are important for relating our semantics to the model and proof theoretic tradition associated with compositional semantics. Below is an example of a record type:

$$\left[\begin{array}{l} \text{REF} \quad : \quad \text{Ind} \\ \text{SIZE} \quad : \quad \text{size}(\text{REF}, \text{MuchBiggerThanMe}) \\ \text{SHAPE} \quad : \quad \text{shape}(\text{REF}, \text{BearShape}) \end{array} \right]$$

A record of this type has to have fields with the same labels as those in the type. (It may also include additional fields not required by the type.) In place of the types which occur to the right of ‘:’ in the record type, the record must contain an object of that type. Here is an example of a record of the above type:

$$\left[\begin{array}{l} \text{REF} \quad = \quad \text{obj123} \\ \text{SIZE} \quad = \quad \text{sizesensorreading85} \\ \text{SHAPE} \quad = \quad \text{shapesensorreading62} \\ \text{COLOUR} = \quad \text{coloursensorreading78} \end{array} \right]$$

Thus, for example, what occurs to the right of the ‘=’ in the REF field of the record is an object of type *Ind*, that is, an individual. Types which are constructed with predicates like *size* and *shape* are sometimes referred to as “types of proof”. The idea is that something of this type would be a proof that a given individual (the first argument) has a certain size or shape (the second argument). One can have different ideas of what kind of objects count as proofs. Here we are assuming that the proof-objects are readings from sensors. This is a second way (in addition to the progression of local resources towards general resources) that our theory interfaces with a statistical non-categorical world. We imagine that the mapping from sensor readings to types involves sampling of analogue data in a way that is not unsimilar to the digitization process involved, for example, in speech recognition. Again we have nothing detailed to say about this at the moment, although we regard it as an important part of our theory that it is able to make a connection between the realm of feature vectors and the realm of model-theoretic semantics.

Types constructed with predicates may also be *dependent*. This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ‘:’ elsewhere in the record type. This means, for example, that in considering whether a record is of the record type, you will need to find a proof that the object which is in the REF-field of the record has the size represented by *MuchBiggerThanMe*. That is, this type depends on the value for the REF-field.

Some of our types will contain *manifest fields* (Coquand et al., 2004) like the REF-field in the following type:

$$\left[\begin{array}{l} \text{REF=obj123} \quad : \quad \text{Ind} \\ \text{SIZE} \quad : \quad \text{size}(\text{REF}, \text{MuchBiggerThanMe}) \\ \text{SHAPE} \quad : \quad \text{shape}(\text{REF}, \text{BearShape}) \end{array} \right]$$

$[\text{REF=obj123:Ind}]$ is a convenient notation for $[\text{REF : Ind}_{\text{obj123}}]$ where $\text{Ind}_{\text{obj123}}$ is a *singleton type*. If $a : T$, then T_a is a singleton type and $b : T_a$ (i.e. b is of type T_a) iff $b = a$. Manifest fields allow us to progressively specify what values are required for the fields in a type.

An important notion in this kind of type theory is that of *subtype*. For example,

$$\left[\begin{array}{l} \text{REF} \quad : \quad \text{Ind} \\ \text{SIZE} \quad : \quad \text{size}(\text{REF}, \text{MuchBiggerThanMe}) \end{array} \right]$$

is a subtype of

$$[\text{REF} \quad : \quad \text{Ind}]$$

as is also

$$[\text{REF=obj123} \quad : \quad \text{Ind}]$$

The subtype relation corresponds to that of *subsumption* in typed feature structures. This gives us the ability to create type hierarchies corresponding to ontologies (in the sense, for example, of OWL). Such ontologies (coded in terms of record types) play an important role in our notion of resources available to an agent. In fact, modelling concepts in terms of record types commits us to a view of concepts which is very closely related to work on ontologies. But our view of the creation of local situation specific and domain related resources in addition to generic resources means that agents have access not to a single generic ontology but also situation specific and domain related ontologies. And, perhaps most important of all, the process of semantic coordination with an interlocutor can involve local *ad hoc* adjustment to an ontology. This plays an important role in characterizing the options open to an agent when confronted with an innovative utterance. We attempt to illustrate this below by working in more detail through a specific example.

3.2 “Panda” as an example of innovative use

We provide an analysis of B ’s utterance in our initial example as a move of offering “panda” as an alternative for “bear”, and as potentially triggering an update on A ’s concepts for “bear” and “panda”.

A: That’s a nice bear

B: Yes, it’s a nice panda

The dialogue move analysis of this example is **offer-form:in-repair**(“panda”, “[it] is a nice _”), or in the generalised format **offer-form**(“panda”, “bear”).

We assume that, before B ’s utterance, A has a single concept of “bear” in a domain called “zoo”, that is, a unique member of the collection $[\text{bear}]_{\text{zoo}}^A$. We represent it in Figure 1. A ’s take on the communicative situation where B ’s utterance takes place (that is, A ’s dialogue information state, much simplified for expository reasons) is shown in Figure 2. This is intended to describe a situation at a zoo, where a bear-shaped object much bigger than A is in focus (FOO here stands for “Focused Object”).

What happens after B ’s utterance? First, we assume that B correctly understands A ’s utterance as offering “panda” as an alternative for “bear”. Now, assuming that B has not observed the word “panda” before, A needs to create a panda-concept $[\text{panda}]_s^A$, local to the communicative situation s resulting from B ’s utterance. Since “panda” has been aligned with “bear”, it is natural to base the new panda concept on the bear concept, associated with the domain. Here A is confronted with a fundamental choice. Should a condition ‘panda(REF)’ be added to the concept in addition to the condition ‘bear(REF)’ making the panda concept be a subtype of the bear concept or should the panda condition replace the bear condition, making panda and bear sisters in the ontology? There is not enough evidence in this simple exchange to determine this.¹ We will choose to replace the bear condition with the panda condition. But there is more that must happen.

A has observed that the use of “panda” in s refers to the focused object obj123 . Following the principle of contrast (Clark and Wong, 2002) which states that “(s)peakers take every difference in form to mark a difference in meaning”, B takes “panda” to have a different meaning than “bear” in some respect other than that it is a panda as opposed to a bear, and looks for something about obj123 which might distinguish it from previously observed bears. For example, the child might decide that it is the colour (black and white) that

¹And indeed many people can reach adulthood, the present authors included, without being sure whether pandas are a kind of bear or not.

REF	:	Ind
PHYS	:	phys-obj(REF)
ANIM	:	animate(REF)
SIZE	:	size(REF, MuchBiggerThanMe)
SHAPE	:	shape(REF, BearShape)
BEAR	:	bear(REF)

Figure 1: A’s “bear” concept in the domain “zoo” before the interaction

DOMAIN	:	zoo															
SHARED	:	<table style="border: none; display: inline-table;"> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">FOO=obj123</td> <td style="padding-right: 10px;">:</td> <td>Ind</td> </tr> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">COM=</td> <td style="padding-right: 10px;">:</td> <td> <table style="border: none; display: inline-table;"> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td>nice(FOO)</td> </tr> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td>bear(FOO)</td> </tr> </table> </td> </tr> <tr> <td style="border-right: 1px solid black; padding-right: 10px;"></td> <td style="padding-right: 10px;">:</td> <td>RecType</td> </tr> </table>	FOO=obj123	:	Ind	COM=	:	<table style="border: none; display: inline-table;"> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td>nice(FOO)</td> </tr> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td>bear(FOO)</td> </tr> </table>	C ₁	:	nice(FOO)	C ₂	:	bear(FOO)		:	RecType
FOO=obj123	:	Ind															
COM=	:	<table style="border: none; display: inline-table;"> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td>nice(FOO)</td> </tr> <tr> <td style="border-right: 1px solid black; padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td>bear(FOO)</td> </tr> </table>	C ₁	:	nice(FOO)	C ₂	:	bear(FOO)									
C ₁	:	nice(FOO)															
C ₂	:	bear(FOO)															
	:	RecType															

Figure 2: A’s take on s

distinguishes it from previously observed bears (which have all been brown)². A now creates a situated interpretation $[\text{panda}]_s^A$ of “panda”, based on $[\text{bear}]_{\text{zoo}}^A$, as shown in Figure 3.

But now if colour is being used to distinguish between bears and pandas in situation s , A should create a refined bear concept for s , namely Figure 4 reflecting the hypothesis that bears are brown. If A is optimistic, possibly influenced by the degree of expertise which A assigns to B (“Mummy knows about zoos”), A might immediately associate the concept in Figure 4 with the zoo domain, that is, make it be a new value for $[\text{bear}]_{\text{zoo}}^A$ and similarly for a dereferenced version of Figure 3, that is a version in which the manifest field is replaced by $[\text{REF} : \text{Ind}]$. Finally, A ’s new take on s is shown in Figure 5; A has accepted that the focused object is a panda.

4 Conclusion

We have sketched an account of how concepts can be updated as a result of language use in interaction. Such processes enable coordination of domain-specific microlanguages, involving a domain-specific grammar and lexicon, an ontology, and a mapping between lexicon and ontology.

There are many mechanisms for semantic coordination, some of which can be described as corrective feedback: clarification requests, explicit corrections, meaning accommodation (observing instances of language use and silently adapting to successful instances) and explicit negotiation. Semantic coordination, in turn, is a kind of language coordination (other kinds include e.g. phonetic co-

ordination). Finally, language coordination coexists with information coordination, the exchanging and sharing of information (agreeing on relevant information and future action; maintaining a shared view on current topics of discussion, relevant questions etc.). Arguably, the main point of language coordination is to enable information coordination.

Semantic coordination happens in dialogue; it is part of language coordination; and it is a prerequisite for information coordination. If we say that a linguistic expression c has meaning only if it is possible to exchange information using c , then semantic coordination is essential to meaning. A linguistic expression c has meaning in a language community when the community members are sufficiently coordinated with respect to the meaning of c to allow them to use c to exchange information. In other words: meaning emerges from a process of semantic coordination in dialogue.

Acknowledgement

This research was supported by The Swedish Bank Tercentenary Foundation Project P2007/0717, Semantic Coordination in Dialogue.

²This account relies on A having a memory of previously observed instances of a concept, in addition to the concept itself (which in the case of “bear” does not contain information about colour).

REF= <i>obj123</i>	:	Ind]
PHYS	:	phys-obj(REF)	
ANIM	:	animate(REF)	
SIZE	:	size(REF, MuchBiggerThanMe)	
SHAPE	:	shape(REF, BearShape)	
COLOUR	:	colour(REF, BlackAndWhite)	
PANDA	:	panda(REF)]

Figure 3: *A*'s situated interpretation of “panda” in situation *s*.

REF	:	Ind]
PHYS	:	phys-obj(REF)	
ANIM	:	animate(REF)	
SIZE	:	size(REF, MuchBiggerThanMe)	
SHAPE	:	shape(REF, BearShape)	
COLOUR	:	colour(REF, Brown)	
BEAR	:	bear(REF)]

Figure 4: *A*'s local “bear” concept after integrating *B*'s utterance

DOMAIN	:	zoo]																
SHARED	:	<table style="border-collapse: collapse; margin-left: 20px;"> <tr> <td style="padding-right: 10px;">FOO=<i>obj123</i></td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">Ind</td> <td style="padding-left: 10px;">]</td> </tr> <tr> <td>COM=</td> <td>:</td> <td> <table style="border-collapse: collapse; margin-left: 20px;"> <tr> <td style="padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">nice(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> <tr> <td style="padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">panda(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> </table> </td> <td style="padding-left: 10px;">]</td> </tr> </table>	FOO= <i>obj123</i>	:	Ind]	COM=	:	<table style="border-collapse: collapse; margin-left: 20px;"> <tr> <td style="padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">nice(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> <tr> <td style="padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">panda(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> </table>	C ₁	:	nice(FOO)]	C ₂	:	panda(FOO)]]]: RecType
FOO= <i>obj123</i>	:	Ind]																
COM=	:	<table style="border-collapse: collapse; margin-left: 20px;"> <tr> <td style="padding-right: 10px;">C₁</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">nice(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> <tr> <td style="padding-right: 10px;">C₂</td> <td style="padding-right: 10px;">:</td> <td style="padding-right: 10px;">panda(FOO)</td> <td style="padding-left: 10px;">]</td> </tr> </table>	C ₁	:	nice(FOO)]	C ₂	:	panda(FOO)]]								
C ₁	:	nice(FOO)]																
C ₂	:	panda(FOO)]																

Figure 5: *A*'s revised take on *s*

References

- S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:482–493.
- H. H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.
- Eve V. Clark and Andrew D. W. Wong. 2002. Pragmatic directions about language use: Offers of words and relations. *Language in Society*, 31:181–212.
- E. V. Clark. 2007. Young children’s uptake of new words in conversation. *Language in Society*, 36:157–82.
- Robin Cooper and Aarne Ranta. 2008. Natural languages as collections of resources. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Relating Dialogue Coordination to Language Variation, Change and Evolution*. College Publications, London.
- Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3:333–362.
- Robin Cooper. 2008. Type theory with records and unification-based grammar. In Fritz Hamm and Stephan Kepser, editors, *Logics for Linguistic Structures*. Mouton de Gruyter.
- Thierry Coquand, Randy Pollack, and Makoto Takeyama. 2004. A logical framework with dependently typed records. *Fundamenta Informaticae*, XX:1–22.
- Simon C. Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.
- P.G.T. Healey. 1997. Expertise or expertese?: The emergence of task-oriented sub-languages. In M.G. Shafto and P. Langley, editors, *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 301–306.
- Staffan Larsson. 2007. Coordinating on ad-hoc semantic systems in dialogue. In *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue*.
- Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–226, April.
- David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Ronnie Smith and Jan Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*. Kluwer Academic Publishers.

A Collaborative Tool for the Computational Modelling of Child Language Acquisition

Kris Jack

CEA LIST, Laboratoire d'ingénierie de la connaissance multimédia multilingue
18 route du Panorama, BP6
Fontenay-aux-Roses, F-92265 France
mrkrisjack@gmail.com

Abstract

A large number of computational language learners have been proposed for modelling the process of child language acquisition. Comparing them, however, can be difficult due to the different assumptions that they make, the diverse test results presented, and the different linguistic behaviours investigated. This paper introduces a toolkit that allows different language learners to be trained, tested and analysed under standardised conditions. The results can be easily compared with one another and with typical child language development to highlight the relative advantages and disadvantages of learners.

1 Introduction

The computational modelling of language acquisition can help in understanding the acquisition process by estimating the problem faced by children and designing algorithms that solve it in a similar way as they do. Many such models have been produced in recent years, tackling various linguistic behaviours. Like in any relatively new domain of research, however, the treatment of one problem often reveals the presence of several more that in turn require new solutions of their own. This has led to the design and implementation of numerous learners that differ in either subtle or fundamental ways. Given such variety, it is not yet clear which kind of model, or combination of models, can best account for the overall behaviour witnessed during child language development.

When surveying the computational language acquisition literature, the relative advantages and

disadvantages of language learners are not always clear. As such, it can be very difficult to compare different learners with one another. The main problem is the lack of standardisation in the field. Language learners are constructed with different underlying assumptions, largely due to the lack of consensus in linguistic theory, are trained using different data (that can vary from miniature languages to full blown natural languages) and are tested using different testing measures (some of which include the '*Looks good to me*' approach).

In this paper, a toolkit for investigating the computational modelling of child language acquisition is proposed. The Language Acquisition Toolkit (LAT) allows researchers to work collaboratively in solving the modelling task, while addressing the problems introduced. It is an attempt to bring the field forward by creating a standardised way for testing and implementing language acquisition learners. The issues addressed in this paper are largely driven by engineering concerns although the choices that are made by the modeller will impact not only on their learner but also on the associated language theory. The driving motivation behind the LAT is that the best way to compare different language learners is to compare the behaviours that they produce. The closer a learner's behaviour is to the behaviour witnessed in children, the better the model.

The LAT is a computational framework that can train, test and analyse the linguistic performance of a computational language learner in order to chart developmental linguistic trajectories. The motivation for the LAT shall first be explored before describing it in detail, discussing its features and considering future directions.

2 Background

The process of modelling child language acquisition is very complex, as many of the first attempts confirmed (Feldman *et al.*, 1990; Suppes, Liang & Bottner, 1991). Rather than modelling the process in entirety, an undoubtedly daunting task, modellers took the simplified approach of focusing upon individual linguistic behaviours, leading to much research into relatively constrained problems such as understanding over- and under-generalisation errors (Plunkett, Sinha, Moller & Strandsby, 1992), single word learning (Regier, 2005), syntactic category acquisition (Redington, Chater & Finch, 1988) and past-tense learning (Rumelhart & McClelland, 1986). While such models have led to valuable insights in the domain, it can be difficult to see how each of them is related to one another given the lack of standardised learning, testing and analysis.

Often, the variety found in computational models reflects the divisions between linguistic theories pertaining to child language acquisition (Kaplan, Oudeyer & Bergen, 2008). Given that linguists remain divided about how children learn language, it is not surprising to find a similar division in the computational modelling community as well. One of the fundamental issues that separates modellers is the kind of data that the learner learns. This can range from the use of plain textual data (Elman, 1993), to grounded sensor-based input (Roy, 2008). Standardising the type of learning data would thus be useful for comparing language learners.

Typical computational models are often tested under different circumstances and using different techniques. For example, while some papers offer a general analysis of the model's behaviour, others focus on particular features, while some test language comprehension, others test language production, and while some consider developmental growth, others consider only the start and end points of training. Although this is often justifiable in the context of the research problem, it makes it difficult to directly compare two models. It would be useful to put all models through the same set of rigorous tests in order to find out how they are similar and how they differ from one another. Such standardised testing will often reveal important differences that may have previously been hidden.

Practically, however, not all models that are described in the literature are made available for

download. As a result, researchers often have to spend time recreating models. This assumes, of course, that the model has been described in enough detail that it can be faithfully recreated. Much time could be saved if such models were available for download, from a common repository, such as the Weka makes machine learning algorithms freely available in a software workbench (<http://www.cs.waikato.ac.nz/~ml/>).

A good language learner should not just solve language learning problems, but should do so in a similar way as is witnessed in children. Based on psycho-linguistic evidence, several linguistic timetables have been derived containing important linguistic milestones (Brown, 1973; Ingram, 1989; Pinker, 1994; Tomasello, 2005). The character of language development is a significant feature in child language acquisition and modellers should be encouraged to model it to better understand the process. A language learner that demonstrates a good use of syntax at the same time as producing its first words is not very realistic. Instead, there should be a prolonged period in which words are learned followed by the emergence of syntax. Unfortunately, a language model can often produce behaviours at unexpected times, signalling a problem with the linguistic theory that it embodies. A standardised approach to analysing the linguistic development of a language learner would be an advantage.

3 The Language Acquisition Toolkit

3.1 Introduction

The Language Acquisition Toolkit (LAT) is a piece of software that allows researchers to develop and test computational language learners within a standardised environment. The LAT's target users are researchers who have basic skills in software development and are comfortable using the programming language Java. It assumes that the language learner operates under the restrictions imposed in the miniature language paradigm (Feldman *et al.*, 1990). The LAT can be obtained from www.langac.com and is available under a GNU public license meaning that the code can be reproduced and modified without obtaining permission.

The LAT is an attempt to standardise the training, testing and analysing of language learners within an open and accessible environment (Figure 1). In training, the language learner observes a

simulated world in which action-based events occur. Both simulated descriptions and visual data are sent to the language learner for analysis. The LAT then tests both the language learner's comprehension and production capacities. Comprehension is tested by sending a description to the language learner and scoring the visual data that are produced. Similarly, production is tested by sending visual data to language learner and scoring the descriptions produced. The LAT then analyses the results obtained from testing and develops data describing the learner's development.

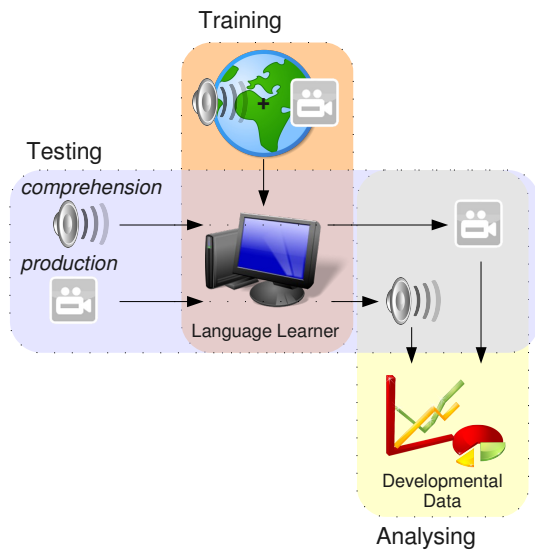


Figure 1: LAT Overview. A language learner is placed in the LAT's simulated world where it learns from simulated audio and visual data. The LAT tests the learner and the results are used to produce data describing its development.

3.2 Training

The LAT can be configured to train different language learners by generating a simulated environment in which action-based events occur. The simulated environment operates within the miniature language acquisition paradigm (Feldman *et al.*, 1990), a simplified simulation of the real-world. A simulation is employed rather than grounding the model in the real-world in order to better control the number and type of problems that are being investigated in a single experiment. While the miniature language paradigm imposes a number of constraints, the proposed simulation contains enough complexity to justify its use.

The learner is trained by watching an event that is simulated in the blocks world in which a number of geometric objects can be found. When an event occurs, a symbolic representation of the description and visual data are generated.

More concretely, an event is the pairing of a simulated description and a action, $e = \langle d, a \rangle$. Events are represented following evidence from child studies. First, it is assumed that the learner can establish a triadic relationship between an object, a speaker and themselves in order to associate a description with an action. This kind of relationship is typically called joint-attention and does not appear in children until around 12 months-old (Tomasello, 1995). As such, the symbolic content present in descriptions and actions are limited to those found in child literature during the first year of life.

An infant's acoustic sensitivity is so attuned that from four-days-old she demonstrates the ability to differentiate between native and non-native speech (Mehler *et al.*, 1988). Such discrimination lies in rhythmic properties that differ over language groups (Dehaene-Lambertz & Houston, 1998; Mehler, Dupoux, Nazzi & Dehaene-Lambertz, 1996) and is likely to be syllable-based since infants detect change in syllable quantity, but not in phoneme quantity over samples of speech (Bijeljic-Babic, Bertoncini & Mehler, 1993). Infants also detect vowel change, a syllable covariant, more readily than consonant change (Bertoncini *et al.*, 1988), further supporting a syllabic base. A description is thus represented as a non-zero length ordered list of syllables in the LAT. Word segmentations are not included as there is no acoustic equivalent of the blank space in written language.

In terms of visual sensitivity, infants can identify objects through retinal and object displacement during motion from four months-old (Kellman, Gleitman & Spelke, 1987), and make relative spacial distinctions between left and right, and above and below, from three to ten months old (Quinn & Schyns, 2003). Infants can also make use of shape and colour to differentiate between objects in the first year of life (Landau, Smith & Jones, 1988). The LAT thus describes the physical properties of objects that inhabit the blocks world (e.g. shape, colour, size and position), referred to as features. An action is defined as a non-zero length ordered list of feature sets, where each feature set is associated with a unique time interval. A set of features describes all objects that can be seen in an event. Note that actions in this terminology do not relate to actions in terms of verbs in natural language, but to a list of descriptions of scenes. Properties such as

push and pull are thus not explicitly represented as symbolic features.

Two types of events can occur in the blocks world: action-based; and descriptive. In the case of an action-based event, an object performs an action while in the case of a descriptive event, objects do not change. As a result, action-based events contain different feature sets, giving the impression of change, while descriptive events contain the same feature sets, indicating no change. The description in an action-based event describes the action while the description in a descriptive event describes an object in the static scene. Objects can perform several actions including moving, flashing, growing, shrinking, appearing, disappearing, destroying another object, hitting another object, pushing another object and pulling another object.

The LAT randomly generate events that can be used as training data. It can create objects, make them perform actions, and describe the events by instantiating appropriate grammar fragments. To encourage the use of standardised sets of training data, a number of sets of data have been randomly generated that each contain 10,000 events. These data have been generated from different parameters (e.g. amount of noise, probability that an object will perform an action in an event, probabilities for each action to occur, number of time intervals for an event, number of colours/shapes/sizes/actions possible) with different language properties (e.g. recursion present/not present, number of rules, language in use).

To provide concrete examples of typical LAT training data, one data set, called the Appearance data set will be presented in detail. The appearance data set is inspired from a study with real participants. Participants sat in front of a computer screen that initially showed a blank white screen. They were asked to describe all changes that were made to the screen in enough detail that a stranger could recreate the scene using only their descriptions. By pressing a key on the keyboard, a new geometric object appeared on the screen and the change was described by the participant. While the addition of an object to a scene appears to be a trivial change, participants produced complex linguistic descriptions that revealed a deep knowledge of their language. For example, descriptions such as “a blue circle appeared to the upper right of the green square at the bottom” and “a red circle appeared between

the four squares making the shape of a cross” (Jack, 2005).

Given the complexity of the language produced, a simplified version the task was constructed in which only the appearance of one object next to another object was considered. By restricting the context, there is less demand for a computational language learner to have a rich semantic representation of scenes. This served as a reasonable starting point from which to conduct the investigation. The actions in the Appearance data set were constructed by randomly generating one object and placing it in the middle of a 3x3 grid scene and then adding a second object, which was also randomly generated, in a different position. Eight colours and shapes were used. Each action was also accompanied by an appropriate description that was generated using a grammar fragment (Figure 2).

$E \rightarrow NP_1 PAR_2$	
$PAR_1 \rightarrow NP_1 PART$	$PAR_2 \rightarrow REL NP_2$
$RELT \rightarrow REL Det_2$	$REL \rightarrow REL_1 REL_2$
$REL_1 \rightarrow a\ bove\ \ be\ low\ \ to\ the\ REL_4$	$REL_2 \rightarrow REL_3\ REL_4$
$REL_3 \rightarrow to\ the\ low\ er\ \ to\ the\ u\ pper$	$REL_4 \rightarrow left\ of\ \ right\ of$
$NP_1 \rightarrow Det_1\ N_{bar}$	$NP_2 \rightarrow Det_2\ N_{bar}$
$N_{bar} \rightarrow SHP\ COL$	
$Det_1 \rightarrow a$	$Det_2 \rightarrow the$
$COL \rightarrow black\ \ blue\ \ grey\ \ green\ \ pink\ \ black\ \ red\ \ ye\ low$	$SHP \rightarrow cir\ cle\ \ cross\ \ dia\ mond\ \ heart\ \ rec\ tang\ le\ \ star\ \ square\ \ tri\ ang\ le$

Figure 2: Miniature Language from Appearance Data Set. All strings are syllable segmented rather than word segmented.

Events from this data set have actions that are described using a 2-frame time interval, where the first set of features describes the state of the scene before the action occurs and the second set of features describes the scene after the action occurs (Figure 3). Note that it is assumed that the learner can identify concepts such as colour, shape and position and that such symbolic information is associated with a particular object. The notion of object-hood, where the first object in the scene is O_1 and the second object is O_2 , is carried across time intervals with O_1 being recog-

nised as the same object before and after an action occurs.

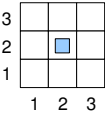
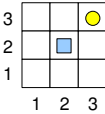
Before action (t=1)	After action (t=2)
	
O ₁ : square O ₁ : blue O ₁ : x2 O ₁ : y2	O ₁ : square O ₂ : circle O ₁ : blue O ₂ : yellow O ₁ : x2 O ₂ : x3 O ₁ : y2 O ₂ : y3
<i>a ye low cir cle to the u pper right of the blue square</i>	

Figure 3: Sample Event from Appearance Data Set. Two time frames represented graphically and as feature sets. The accompanying syllable segmented description of the event is also shown.

The remaining data sets contain more complex events in which more actions and richer miniature languages are employed. Actions are randomly generated, with respect to the constraints imposed on the data set (e.g. number of colours, shapes, and actions) and appropriate descriptions are generated. These descriptions are produced by following a heuristic that minimises the number of syllables that can appear in a single description. This reduces the production of unnatural sentences. For example, take the case where an object appears in a scene amongst 10 other objects. A description could be generated to describe the action with respect to one other object, two other objects or as many as 10 other objects. While such descriptions are all valid, many of them would sound unnatural if employed. The algorithm selects descriptions by favouring those that have fewer syllables. A parser is then employed that eliminates invalid descriptions that can be misinterpreted. By making a parsimonious use of syllables, more natural descriptions tend to be produced. More abstract language can also be found such as the use of the word 'bullying' to describe pushing, pulling and hitting.

3.3 Testing

The LAT monitors the linguistic development of a language learner by testing its comprehension and production capacities. The learner's comprehension and production are tested at every round of training.

For each set of training data, there is an associated set of testing data, ensuring a standardised test procedure for language learners. Test data is produced using grammar rules for producing descriptions and heuristics for producing actions. The tests are constructed to reflect the properties found in the training data's miniature language. As such, the learner is only tested on the kind of descriptions and actions that it has the opportunity to learn through observing events. Concretely, a testing set is a set of events where each event relates one or more descriptions to one or more actions. The set of testing data associated with the Appearance training data set can be used to test the learner's vocabulary, certain multi-word combinations and full sentences. Using the terminology found in Appearance's grammar fragment (Figure 2), the LAT tests for the comprehension of shapes (SHP), colours (COL), objects (N_{bar}), indefinite objects (NP₁), definite objects (NP₂) and events (E).

In testing the learner's comprehension, the LAT sends a description as input and receives a set of actions as output. The output is automatically scored by comparing it with the expected output that is associated with the description. Actions are compared based on the feature values that are relevant to the given description. Given the description "a ye low cir cle to the u pper right of the blue square" (Figure 3), the colours, shapes and relative positions of the objects are relevant whereas their exact positions are not. The LAT equally accepts a yellow circle that appears higher or further right than its idealised position with respect to the blue square, as long as the relative positions remain correct.

Borrowing from research in child language acquisition studies, four kinds of incorrect responses are identified: over-extended; under-extended; mismatched; or incorrect. For example, the meaning of the description "square" is under-extended if the learner only uses it to refer to red squares, blue squares and green squares, but not to squares of other colours. Similarly, the meaning of the description "red square" is over-extended if it refers to red squares, blue squares and red circles. A mismatch is found if the description "square" is used to refer to objects other than squares, for examples circles and triangles, but never to squares themselves. Results that deviate from these cases are simply considered incorrect.

The LAT can score both single words and phrases based on these categories.

In addition, the output produced by the learner can also be described using the standard information retrieval measures of precision, recall, and the e-measure which is a weighted combination of the two former values (van Rijsbergen, 1979).

The process of testing the learner's production is similar to that of testing comprehension. Rather than the LAT sending a description as input, however, it sends an action. The learner then produces a set of descriptions as output. Results from production are scored using the same principles as applied during comprehension. That is, the learner's output is compared to the expected output and it is scored as either correct, over-extended, under-extended, mismatched or incorrect.

3.4 Analysing

Both the comprehension and production results that are produced from testing are used to evaluate the learner's linguistic stage of development. Several types of analysis have been designed to ease learner comparisons: round-based; trial-based; and learner-based. Round-based analyses analyse the results produced from a single round of testing. Trial-based analyses take round-based statistics and compare them with previous rounds in order to find behavioural trends in the data. Finally, learner-based analyses compare trial-based data for several trials in order to extract general behavioural trends. By performing analyses at all three levels of detail, a more complete account of the learner's behaviour is produced.

The LAT is currently able to perform a number of round-based analyses that are often found in the literature: summary of test results in terms of correct results and errors; chart the linguistic generativity of the learner; and present evidence of syntactic activity.

Round-based analyses produce results that are then used to determine the model's stage of linguistic development using data from child language studies: pre-linguistic; holophrastic; early multi-word; late multi-word; and abstract stages.

A number of trial-based analyses are performed using these data, in order to identify particular linguistic behaviours: linguistic development; vocabulary acquisition; comprehension/production imbalance. With the creation of a linguistic development timetable, all data can also

be presented in terms of stages. For example, the number of words that are correctly comprehended and the rate of vocabulary acquisition can be shown by stage.

Model-based analyses can be performed when the results from several trials are available. Each of the results, such as the rate of vocabulary acquisition during a stage, are compared across trials to identify general behavioural trends.

The LAT thus offers a standardised platform for training, testing and analysing language models. The results from all analyses can be automatically compared to determine the differences between learners and which learner best fits child language data.

4 Discussion

The LAT is a freely available tool that offers a standardised environment from which language modellers can develop their language learners. It is an attempt to advance the domain by offering a platform where common goals can be focussed upon in a collaborative environment. It aims to standardise the training, testing and analysing of language learners by understanding the needs of language modellers through collaboration.

By using the LAT, the language modeller accepts the need to work with standardised training data. Such standardisation is widespread in computational linguistics. For example, in the field of automatic text classification, there are several databases of pre-classified documents (e.g. Reuters-21578, Reuters Corpus Volume 1 and 20 Newsgroups) that researchers can use to evaluate different algorithms and to compare their results. The LAT offers different sets of training data that are constrained by principles of the miniature language paradigm. In using such data, the modelling task differs from the task that a child faces in a number of ways. In particular, the learning problem is simplified in that the real-world contains many more objects and that natural language has far more linguistic structures and words than the language fragments. It is for these reasons, however, that such a paradigm is attractive. Many language learning problems can be effectively investigated by first simplifying the problem and then developing solutions. When such problems in the miniature language paradigm have been adequately solved, it is envisaged that the LAT can be grounded in a real

environment where vast volumes of data are available for processing.

The results from learning can then be tested using a standardised set of tests. The learner is treated as a black box, meaning that the LAT evaluates its output alone without entering into its inner workings. This helps to keep the LAT's functionality independent from the learner by focussing on the way in which it behaves rather than how it produces particular behaviours, similar to the relationship found between the linguist and child in the real world. By testing both comprehension and production on a large set of descriptions and actions, a complete picture of the learner's linguistic state can be derived. The LAT also checks for language errors such as over-extensions, under-extensions and mismatches. Individual results are made available to the researcher in a tabular format as well as providing overall recall, precision and e-measure scores.

By standardising the test results, different language learners can be easily compared with one another. The LAT can analyse these results to discover behavioural trends in the data which can be used in further comparisons. It is also interesting to note that the LAT makes an attempt to compare the behaviour produced by a language learner with that of children. Inspired by child language development timetables, a set of milestones has been derived that are used to characterise the learner's behaviour in terms of stages. The LAT attempts to encourage researchers to consider the developmental behaviour of their language learners over time.

It is important to note that the LAT is a work in progress. This disclaimer is likely to remain true for many years. Developing a gold standard is a difficult task and one that risks to evolve over time. The LAT should be regarded as a proposal for standardisation. Being a collaborative project, any contributor can challenge this proposal by offering their own solutions. Contributors are encouraged to create their own data and algorithms and to upload them to the LAT. A gold standard can only emerge from the selections that are made by other modellers, who vote by using certain data and algorithms in their own modelling tasks. In this sense, the proposed instantiation of the LAT described in this article is less important than the idea behind the LAT itself.

5 Future Considerations

In designing the LAT, it quickly became clear that the task was not straight-forward. Designing a tool that can make useful and standardised comparisons between language learners is a complex task. A balancing act between not excluding certain types of learners and creating a constrained, manageable environment is not without its difficulties. As such, it is worth considering future developments for the LAT. While still in a preliminary state of development, it is hoped that a collaborative approach to the task will allow it to be steered in the directions that are best adapted to its potential users. A number of these directions are now considered.

The miniature language paradigm is at the heart of the LAT. This language can be extended to include more complex linguistic constructions and a larger vocabulary. It is suggested that a systematic approach is followed in which the learning task is made progressively complex by adding linguistic features that tend to be witnessed in children during development. It seems reasonable to follow a longitudinal approach to development. Contributors are also encouraged to create and submit new training data sets in order to explore how complex a miniature language can become.

The type of information that is available to the learner could also be changed. At present, the descriptions lack acoustic information such as tone. Such data is indispensable in investigating certain languages such as Mandarin and Swahili. Similarly, the symbolic representations of visual objects can be refined to better represent reality. Colours can be represented by RGB values rather than linguistically-related symbols, as it is unlikely that children start with such pre-defined semantic categories from the outset of learning.

It is also worthwhile considering more complex testing and analysis algorithms. It is likely that they will be developed in step with new linguistic phenomena that are investigated, building a useful catalogue of tools. In addition, it may be useful to develop learner-dependant analysis tools in order to demonstrate how the inner workings are related to the outward behaviour.

Finally, it is hoped that the LAT will become a useful resource not just for modellers who are comfortable with coding but also non-programmers. They should be able to implement and experiment with different kinds of models with the

flexibility of looking at different aspects of acquisition under different settings and with different types of data. They can then inform language modellers directly about how particular language models perform well and poorly in certain cases. The collaborative aspect of the LAT encourages not just programmers to share their code, but for everyone to share their ideas.

6 Conclusion

This article proposes a tool that facilitates the consolidation of research into the computational modelling of child language acquisition under the miniature language paradigm. The workshop is being used to launch a first version of the LAT, that is hoped to help language modellers and child language experts to communicate and share their knowledge.

7 Acknowledgements

This research was supported by the Jean-Luc Lagardère Foundation (<http://www.fondation-jeanlucagardere.com>).

Many thanks to the anonymous reviewers for their constructive comments.

References

- Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L. & Mehler, J. (1988). An investigation of Young infants' perceptual representations of speech sounds. *Journal of experimental psychology*, 117, pp. 21-33.
- Bijeljac-Babic, R., Bertoncini, J. & Mehler, J. (1993). How Do 4-Day-Old Infants Categorize Multisyllabic Utterances?. *Developmental Psychology*, 29, pp. 711-721.
- Brown, R. (1973). *A First Language: The Early Stages*. Harvard University Press.
- Dehaene-Lambertz, G. & Houston, D. (1998). Faster Orientation Latencies Toward Native Language in Two-Month-Old Infants. *Language and Speech*, 41, pp. 21-43.
- Elman, J. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48, pp. 71-99.
- Feldman, J., Lakoff, G., Stolcke, A. & Weber, S. (1990,). *Miniature Language Acquisition: A Touchstone for Cognitive Science*.
- Ingram, D. (1989). *First Language Acquisition: Method, Description and Explanation*. Cambridge University Press.
- Jack, K. (2005,). *Introducing a Scene Building Game to Model Early First Language Acquisition*, CLUK, Manchester, England.
- Kaplan, F., Oudeyer, P. & Bergen, B. (2008). Computational models in the debate over language learnability. *Infant and Child Development*, 17, pp. 55-80.
- Kellman, P., Gleitman, H. & Spelke, E. (1987). Object and observer motion in the perception of objects by infants. *Journal of experimental psychology. Human perception and performance*, 13, pp. 586-593.
- Landau, B., Smith, L. & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, pp. 299-321.
- Mehler, J., Dupoux, T., Nazzi, T. & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. Lawrence Erlbaum.
- Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J. & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, pp. 143-178.
- Pinker, S. (1994). *The Language Instinct*. William Morrow.
- Plunkett, K., Sinha, C., Moller, M. & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, 4, pp. 293-312.
- Quinn, P. & Schyns, P. (2003). What goes up may come down: perceptual process and knowledge access in the organization of complex visual patterns by young infants. *Cognitive Science*, 27, pp. 923-935.
- Redington, M., Chater, N. & Finch, S. (1988). Distributional Information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, pp. 425-469.
- Regier, T. (2005). The emergence of words : Attentional learning in form and meaning. *Cognitive Science*, 29, pp. 819-865.
- Roy, D. (2008). A mechanistic model of three facets of meaning. In M. D. Vega, G. Glennberg & G. Graesser (Eds.), *Symbols and Embodiment*. Oxford University Press.
- Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition*. MIT Press. pp. 216-271.
- Suppes, P., Liang, L. & Bottner, M. (1991). Complexity Issues in Robotic Machine Learning of Natural Language. In L. Lam & V. Naroditsky (Eds.), *Modeling Complex Phenomena*. Springer Verlag.
- Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development*. Erlbaum.
- Tomasello, M. (2005). *Constructing a Language : A Usage-Based Theory of Language Acquisition*. Harvard University Press.
- van Rijsbergen, C. J. (1979). *Information Retrieval*. London, Butterworths.

What's in a Message?

Stergos D. Afantenos and Nicolas Hernandez

LINA, (UMR CNRS 6241)

Université de Nantes, France

stergos.afantenos@univ-nantes.fr

nicolas.hernandez@univ-nantes.fr

Abstract

In this paper we present the first step in a larger series of experiments for the induction of predicate/argument structures. The structures that we are inducing are very similar to the conceptual structures that are used in Frame Semantics (such as FrameNet). Those structures are called messages and they were previously used in the context of a multi-document summarization system of evolving events. The series of experiments that we are proposing are essentially composed from two stages. In the first stage we are trying to extract a representative vocabulary of words. This vocabulary is later used in the second stage, during which we apply to it various clustering approaches in order to identify the clusters of predicates and arguments—or frames and semantic roles, to use the jargon of Frame Semantics. This paper presents in detail and evaluates the first stage.

1 Introduction

Take a sentence, any sentence for that matter; step back for a while and try to perceive that sentence in its most abstract form. What you will notice is that once you try to abstract away sentences, several regularities between them will start to emerge. To start with, there is almost always an *action* that is performed.¹ Then, there is most of the times an *agent* that is performing this action and a *patient* or a *benefactor* that is receiving this action, and it could be the case that this action is performed with the aid of a certain *instrument*. In other words, within a sentence—and in respect to its action-denoting word, or predicate in linguistic terms—there will be several entities that are associated with the predicate, playing each time a specific *semantic role*.

The notion of semantic roles can be traced back to Fillmore's (1976) theory of Frame Semantics. According to this theory then, a *frame* is a conceptual structure which tries to describe a stereotypical situation, event or object along with its participants and props. Each frame takes a name (*e.g.* COMMERCIAL TRANSACTION) and contains a list of *Lexical Units (LUs)* which

¹In linguistic terms, an action-denoting word is also known as a *predicate*.

actually evoke this frame. An LU is nothing else than a specific word or a specific meaning of a word in the case of polysemous words. To continue the previous example, some LUs that evoke the frame of COMMERCIAL TRANSACTION could be the verbs *buy*, *sell*, etc. Finally, the frames contain several frame elements or *semantic roles* which actually denote the abstract conceptual entities that are involved with the particular frame.

Research in semantic roles can be distinguished into two major branches. The first branch of research consists in *defining* an ontology of semantic roles, the frames in which the semantic roles are found as well as defining the LUs that evoke those frames. The second branch of research, on the other hand, *stipulates* the existence of a set of frames, including semantic roles and LUs; its goal then, is the creation of an algorithm that given such a set of frames containing the semantic roles, will be able to label the appropriate portions of a sentence with the corresponding semantic roles. This second branch of research is known as *semantic role labeling*.

Most of the research concerning the definition of the semantic roles has been carried out by linguists who are *manually* examining a certain amount of frames before finally defining the semantic roles and the frames that contain those semantic roles. Two such projects that are widely known are the FrameNet (Baker et al., 1998; Ruppenhofer et al., 2006) and PropBank/NomBank² (Palmer et al., 2005; Meyers et al., 2004). Due to the fact that the aforementioned projects are accompanied by a large amount of annotated data, computer scientists have started creating algorithms, mostly based on statistics (Gildea and Jurafsky, 2002; Xue, 2008) in order to automatically label the semantic roles in a sentence. Those algorithms take as input the frame that

²We would like to note here that although the two approaches (FrameNet and PropBank/NomBank) share many common elements, they have several differences as well. Two major differences, for example, are the fact that the Linguistic Units (FrameNet) are referred to as Relations (PropBank/NomBank), and that for the definition of the semantic roles in the case of PropBank/NomBank there is no reference ontology. A detailed analysis of the differences between FrameNet and PropBank/NomBank would be out of the scope of this paper.

contains the roles as well as the predicate³ of the sentence.

Despite the fact that during the last years we have seen an increasing interest concerning semantic role labeling,⁴ we have not seen many advancements concerning the issue of *automatically inducing* semantic roles from raw textual corpora. Such a process of induction would involve, firstly the identification of the words that would serve as predicates and secondly the creation of the appropriate clusters of word sequences, within the limits of a sentence, that behave similarly in relation to the given predicates. Although those clusters of word sequences could not actually be said to serve in themselves as the semantic roles,⁵ they can nevertheless be viewed as containing characteristic word sequences of specific semantic roles. The last point has the implication that if one is looking for a human intuitive naming of the semantic role that is implied by the cluster then one should look elsewhere. This is actually reminiscent of the approach that is carried out by PropBank/NomBank in which each semantic role is labeled as *Arg1* through *Arg5* with the semantics given aside in a human readable natural language sentence.

Our goal in this paper is to contribute to the research problem of frame induction, that is of the creation of frames, including their associated semantic roles, given as input only a set of textual documents. More specifically we propose a general methodology to accomplish this task, and we test its first stage which includes the use of corpus statistics for the creation of a subset of words, from the initial universe of initial words that are present in the corpus. This subset will later be used for the identification of the predicates as well as the semantic roles. Knowing that the problem of frame induction is very difficult in the general case, we limit ourselves to a specific genre and domain trying to exploit the characteristics that exist in that domain. The domain that we have chosen is that of the terroristic incidents which involve hostages. Nevertheless, the same methodology could be applied to other domains.

The rest of the paper is structured as follows. In section 2 we describe the data on which we have applied our methodology, which itself is described in detail in section 3. Section 4 describes the actual experiments that we have performed and the results obtained, while a discussion of those results follows in section 5. Finally, section 6 contains a description of the related work while we present our future work and conclusions in section 7.

³In the case of FrameNet the predicate corresponds to a “Linguistic Unit”, while in the case of PropBank/NomBank it corresponds to what is named “Relation”.

⁴Cf, for example, the August 2008 issue of the journal *Computational Linguistics* (34:2).

⁵At least as the notion of semantic roles is proposed and used by FrameNet.

2 The Annotated Data

The annotated data that we have used in order to perform our experiments come from a previous work on automatic multi-document summarization of events that evolve through time (Afantenos et al., 2008; Afantenos et al., 2005; Afantenos et al., 2004). The methodology that is followed is based on the identification of similarities and differences—between documents that describe the evolution of an event—synchronically as well as diachronically. In order to do so, the notion of *Synchronic and Diachronic* cross document Relations (SDRs),⁶ was introduced. SDRs connect not the documents themselves but some semantic structures that were called *messages*. The connection of the messages with the SDRs resulted in the creation of a semantic graph that was then fed to a Natural Language Generation (NLG) system in order to produce the final summary. Although the notion of messages was originally inspired by the notion of messages as used in the area of NLG, for example during the stage of *Content Determination* as described in (Reiter and Dale, 1997), and in general they do follow the spirit of the initial definition by Reiter & Dale, in the following section we would like to make it clear what the notion of messages represents for us. In the rest of the paper, when we refer to the notion of messages, it will be in the context of the discussion that follows.

2.1 Messages

The intuition behind messages, is the fact that during the evolution of an event we have several activities that take place and each activity is further decomposed into a series of *actions*. Messages were created in order to capture this abstract notion of actions. Of course, actions usually implicate several entities. In this case, entities were represented with the aid of a domain ontology. Thus, in more formal terms a message *m* can be defined as follows:

$$m = \text{message_type}(\text{arg}_1, \dots, \text{arg}_n) \\ \text{where } \text{arg}_i \in \text{Topic Ontology}, i \in \{1, \dots, n\}$$

In order to give a simple example, let us take for instance the case of the hijacking of an airplane by terrorists. In such a case, we are interested in knowing if the airplane has arrived to its destination, or even to another place. This action can be captured by a message of type `arrive` whose arguments can be the entity that arrives (the airplane in our case, or a vehicle, in general) and the location that it arrives. The specifications of such a message can be expressed as follows:

⁶Although a full analysis of the notion of Synchronic and Diachronic Relations is out of the scope of this paper, we would like simply to mention that the premises on which those relations are defined are similar to the ones which govern the notion of *Rhetorical Structure Relations* in Rhetorical Structure Theory (RST) (Taboada and Mann, 2006), with the difference that in the case of SDRs the relations hold across documents, while in the case of RSTs the relation hold inside a document.

```
arrive (what, place)
  what : Vehicle
  place : Location
```

The concepts `Vehicle` and `Location` belong to the ontology of the topic; the concept `Airplane` is a sub-concept of the `Vehicle`. A sentence that might instantiate this message is the following:

The Boeing 747 arrived at the airport of Stanstend.

The above sentence instantiates the following message:

```
arrive ("Boeing 747", "airport of
Stanstend")
```

The domain which was chosen was that of terroristic incidents that involve hostages. An empirical study, by three people, of 163 journalistic articles—written in Greek—that fell in the above category, resulted in the definition of 48 different message types that represent the most important information in the domain. At this point we would like to stress that what we mean by “most important information” is the information that one would normally expect to see in a typical summary of such kinds of events. Some of the messages that have been created are shown in Table 1; figure 1 provides full specifications for two messages.

free	explode
kill	kidnap
enter	arrest
negotiate	encircle
escape_from	block_the_way
give_deadline	

Table 1: Some of the message types defined.

negotiate (who, with_whom, about) who : Person with_whom : Person about : Activity
free (who, whom, from) who : Person whom : Person from : Place ∨ Vehicle

Figure 1: An example of message specifications

Although in an abstract way the notion of messages, as presented in this paper approaches the notion of frame semantics—after all, both messages and frame semantics are concerned with “who did what, to whom, when, where and how”—it is our hope that our approach could ultimately be used for the problem of frame induction. Nevertheless, the two structures have several points in which they differ. In the following section we would like to clarify those points in which the two differ.

2.2 How Messages differ from Frame Semantics

As it might have been evident until now, the notions of messages and frame semantics are quite similar, at least from an abstract point of view. In practical terms though, the two notions exhibit several differences.

To start with, the notion of messages has been used until now only in the context of automatic text summarization of multiple documents. Thus, the aim of messages is to capture the *essential information* that one would expect to see in a typical summary of this domain.⁷ In contrast, semantic roles and the frames in which they exist do not have this limitation.

Another differentiating characteristic of frame semantics and messages is the fact that semantic roles always get instantiated within the boundaries of the sentence in which the predicate exists. By contrast, in messages although in the vast majority of the cases there is a one-to-one mapping from sentences to messages, in some of the cases the arguments of a message, which correspond to the semantic roles, are found in neighboring sentences. The overwhelming majority of those cases (which in any case were but a few) concern *referring expressions*. Due to the nature of the machine learning experiments that were performed, the actual entities were annotated as arguments of the messages, instead of the referring expressions that might exist in the sentence in which the message’s predicate resided.

A final difference that exists between messages and frame semantics is the fact that messages were meant to exist within a certain domain, while the definition of semantic roles is usually independent of a domain.⁸

3 The Approach Followed

A schematic representation of our approach is shown in Figure 2. As it can be seen from this figure, our approach comprises two stages. The first stage concerns the creation of a lexicon which will contain as most as possible—and, of course, as accurately as possible—candidates that are characteristic either of the predicates (message types) or of the semantic roles (arguments of the messages). This stage can be thought of as a filtering stage. The second stage involves the use of unsupervised clustering techniques in order to create the final clusters of words that are characteristic either of the predicates or of the semantic roles that are asso-

⁷In this sense then, the notion of messages is reminiscent of Schank & Abelson’s (1977) notion of *scripts*, with the difference that messages are not meant to exist inside a structure similar to Schank & Abelson’s “scenario”. We would like also to note that the notion of messages shares certain similarities with the notion of *templates* of Information Extraction, as those structures are used in conferences such as MUC. Incidentally, it is not by chance that the “M” in MUC stands for Message (Understanding Conference).

⁸We would like to note at this point that this does not exclude of course the fact that the notion of messages could be used in a more general, domain independent way. Nevertheless, the notion of messages has for the moment been applied in two specific domains (Afantenos et al., 2008).

ciated with those predicates. The focus of this paper is on the first stage.

As we have said, our aim in this paper is the use of statistical measures in order to extract from a given corpus a set of words that are most *characteristic* of the messages that exist in this corpus. In the context of this paper, a word will be considered as being characteristic of a message if this word is employed in a sentence that has been annotated with that message. If a particular word does not appear in any message annotated sentence, then this word will not be considered as being characteristic of this message. In more formal terms then, we can define our task as follows. If by \mathcal{U} we designate the set of all the words that exist in our corpus, then we are looking for a set \mathcal{M} such that:

$$\begin{aligned} \mathcal{M} \subset \mathcal{U} \quad \wedge \\ w \in \mathcal{M} \Leftrightarrow m \text{ appears at least once} \\ \text{in a message instance} \quad (1) \end{aligned}$$

In order to extract the set \mathcal{M} we have employed the following four statistical measures:

Collection Frequency: The set that results from the union of the $n\%$ most frequent words that appear in the corpus.

Document Frequency: The set that results from the union of the $n\%$ most frequent words of each document in the corpus.

tf.idf: For each word in the corpus we calculate its *tf.idf*. Then we create a set which is the union of words with the highest $n\%$ *tf.idf* score in each document.

Inter-document Frequency: A word has inter-document frequency n if it appears in at least n documents in the corpus. The set with inter-document frequency n is the set that results from the union of all the words that have inter-document frequency n .

As we have previously said in this paper, our goal is the exploitation of the characteristic vocabulary that exists in a specific genre and domain in order to ultimately achieve our goal of message induction, something which justifies the use of the above statistical measures. The first three measures are known to be used in context of Information retrieval to capture topical informations. The latter measure has been proposed by (Hernandez and Grau, 2003) in order to extract rhetorical indicator phrases from a genre dependant corpus.

In order to calculate the aforementioned statistics, and create the appropriate set of words, we ignored all the stop-words. In addition we worked only with the *verbs* and *nouns*. The intuition behind this decision lies in the fact that the created set will later be used for the identification of the predicates and the induction of the

semantic roles. As Gildea & Jurafsky (2002)—among others—have mentioned, predicates, or action denoting words, are mostly represented by verbs or nouns.⁹ Thus, in this series of experiments we are mostly focusing in the extraction of a set of words that approaches the set that is obtained by the union of all the verbs and nouns found in the annotated sentences.

4 Experiments and Results

The corpus that we have consists of 163 journalistic articles which describe the evolution of five different terroristic incidents that involved hostages. The corpus was initially used in the context of training a multi-document summarization system. Out of the 3,027 sentences that the corpus contains, about one third (1,017 sentences) were annotated with the 48 message types that were mentioned in section 2.1.

Number of Documents:	163
Number of Token:	71,888
Number of Sentences:	3,027
Annotated Sentences (messages):	1,017
Distinct Verbs and Nouns in the Corpus:	7,185
Distinct Verbs and Nouns in the Messages:	2,426

Table 2: Corpus Statistics.

The corpus contained 7,185 distinct verbs and nouns, which actually constitute the \mathcal{U} of the formula (1) above. Out of those 7,185 distinct verbs and nouns 2,426 appear in the sentences that have been annotated with the messages. Our goal was to create this set that approached as much as possible to the set of 2,426 distinct verbs and nouns that are found in the messages.

Using the four different statistical measures presented in the previous section, we tried to reconstruct that set. In order to understand how the statistical measures behaved, we varied for each one of them the value of the threshold used. For each statistical measure used, the threshold represents something different. For the Collection Frequency measure the threshold represents the $n\%$ most frequent words that appear in the corpus. For the Document Frequency it represents the $n\%$ most frequent words that appear in each document separately. For *tf.idf* it represents the words with the highest $n\%$ *tf.idf* score in each document. Finally for the Inter-document Frequency the threshold represents the verbs and nouns that appear in at least n documents. Since for the first three measures the threshold represents a percentage, we varied it from 1 to 100 in order to study how this measure behaves. For the case of the Inter-document Frequency, we varied the threshold from 1 to 73 which represents the maximum number of documents in which a word appeared.

In order to measure the performance of the statistical measures employed, we used four different evaluation measures, often employed in the information retrieval

⁹In some rare cases predicates can be represented by adjectives as well.

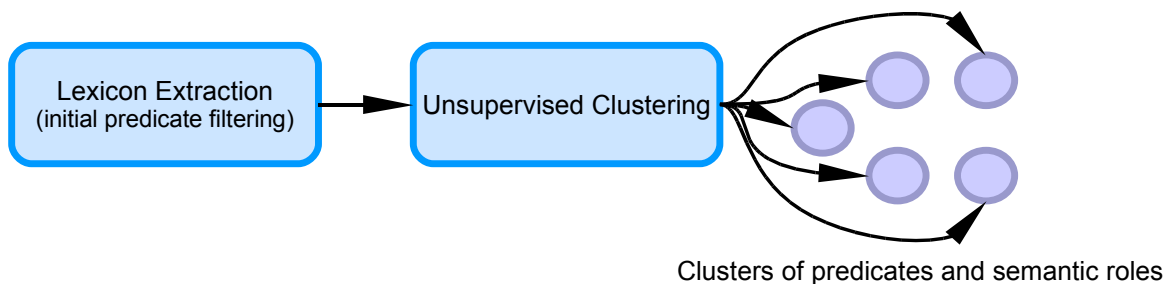


Figure 2: Two different stages in the process of predicate clustering

field. Those measures are the *Precision*, *Recall*, *F-measure* and *Fallout*. Precision represents the percentage of the correctly obtained verbs and nouns over the total number of obtained verbs and nouns. Recall represents the percentage of the obtained verbs and nouns over the target set of verbs and nouns. The F-measure is the harmonic mean of Precision and Recall. Finally, fallout represents the number of verbs and nouns that were wrongly classified by the statistical measures as belonging to a message, over the total number of verbs and nouns that do not belong to a message. In an ideal situation one expects a very high precision and recall (and by consequence F-measure) and a very low Fallout.

The obtained graphs that combine the evaluation results for the four statistical measures presented in section 3 are shown in Figures 3 through 6. A first remark that we can make in respect to those graphs is that concerning the collection frequency, document frequency and tf.idf measures, for small threshold numbers we have more or less high precision values while the recall and fallout values are low. This implies that for smaller threshold values the obtained sets are rather small, in relation to \mathcal{M} (and by consequence to \mathcal{U} as well). As the threshold increases we have the opposite situation, that is the precision falls while the recall and the fallout increases, implying that we get much bigger sets of verbs and nouns.

In terms of absolute numbers now, the best F-measure is given by the Collection Frequency measure with a threshold value of 46%. In other words, the best results—in terms of F-measure—is given by the union of the 46% most frequent verbs and nouns that appear in the corpus. For this threshold the Precision is 54.14%, the Recall is 72.18% and the F-measure is 61.87%. This high F-measure though comes at a certain cost since the Fallout is at 31.16%. This implies that although we get a rather satisfying score in terms of precision and recall, the number of false positives that we get is rather high in relation to our universe. As we have earlier said, a motivating factor of this paper is the automatic induction of the structures that we have called messages; the extracted lexicon of verbs and messages will later be used by an unsupervised clustering algorithm in order to create the classes of

words which will correspond to the message types. For this reason, although we prefer to have an F-measure as high as possible, we also want to have a fallout measure as low as possible, so that the number of false positives will not perturb the clustering algorithm.

If, on the other hand, we examine the relation between the F-measure and Fallout, we notice that for the Inter-document Frequency with a threshold value of 4 we obtain a Precision of 71.60%, a recall of 43.86% and an F-measure of 54.40%. Most importantly though we get a fallout measure of 8.86% which implies that the percentage of wrongly classified verbs and nouns compose a small percentage of the total universe of verbs and nouns. This combination of high F-measure and very low Fallout is very important for later stages during the process of message induction.

5 Discussion

As we have claimed in the introduction of this paper, although we have applied our series of experiments in a single domain, that of terroristic incidents which involve hostages, we believe that the proposed procedure can be viewed as a “general” one. In the section we would like to clarify what exactly we mean by this statement.

In order to proceed, we would like to suggest that one can view two different kinds of generalization for the proposed procedure:

1. The proposed procedure is a general one in the sense that it can be applied in a large corpus of *heterogeneous* documents incorporating various domains and genres, in order to yield “general”, *i.e.* domain-independent, frames that can later be used for any kind of domain.
2. The proposed procedure is a general one in the sense that it can be used in any kind of domain without any modifications. In contrast with the first point, in this case the documents to which the proposed procedure will be applied ought to be *homogeneous* and rather representative of the domain. The induced frames will not be general ones, but instead will be domain dependent ones.

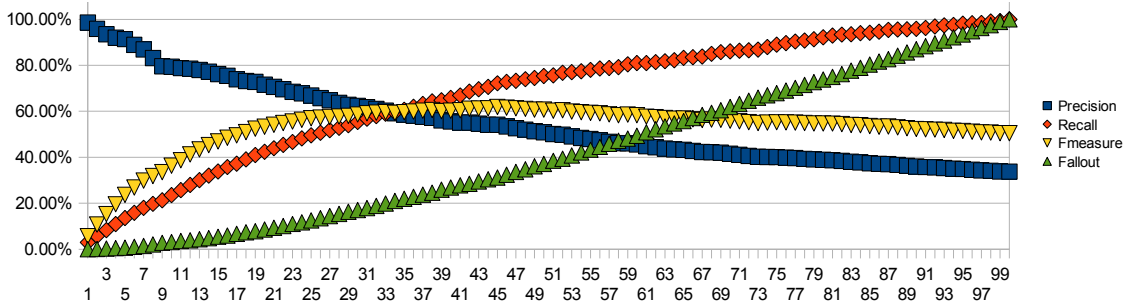


Figure 3: Collection Frequency statistics

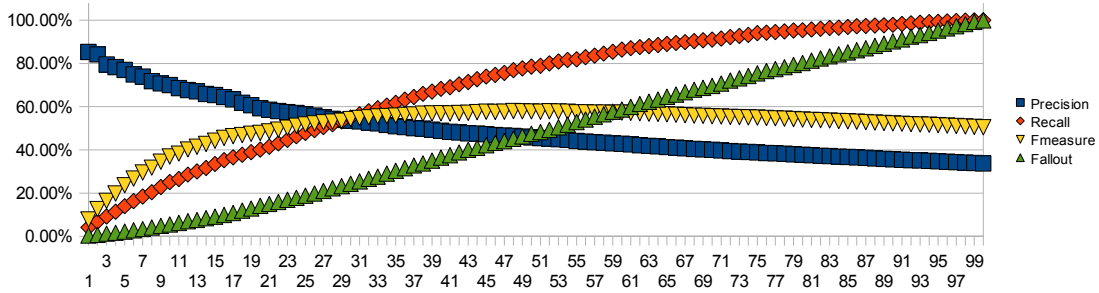


Figure 4: Document Frequency statistics

Given the above two definitions of generality, we could say that the procedure proposed in this paper falls rather in the second category than in the first one. Ignoring for the moment the second stage of the procedure—clustering of word sequences characteristic of specific semantic roles—and focusing on the actual work described in this paper, that is the use of statistical methods for the identification of candidate predicates, it becomes clear that the use of an heterogeneous, non-balanced corpus is prone to skewing the results. By consequence, we believe that the proposed procedure is general in the sense that we can use it for any kind of domain which is described by an homogeneous corpus of documents.

6 Related Work

Teufel and Moens (2002) and Saggion and Lapalme (2002) have shown that templates based on domain concepts and relations descriptions can be used for the task of automatic text summarization. The drawback of their work is that they rely on manual acquisition of lexical resources and semantic classes' definition. Consequently, they do not avoid the time-consuming task of elaborating linguistic resources. It is actually for this kind of reason—that is, the laborious manual work—that automatic induction of various structures is a recurrent theme in different research areas of Natural Language Processing.

An example of an inductive Information Extraction algorithm is the one presented by Fabio Ciravegna

(2001). The algorithm is called $(LP)^2$. The goal of the algorithm is to induce several symbolic rules given as input previous SGML tagged information by the user. The induced rules will later be applied in new texts in order to tag it with the appropriate SGML tags. The induced rules by $(LP)^2$ fall into two distinct categories. In the first we have a bottom up procedure which generalizes the tag instances found in the training corpus which uses shallow NLP knowledge. A second set of rules is also created which have a corrective character; that is, the application of this second set of rules aims at correcting several of the mistakes that are performed by the first set of rules.

On the other hand several researchers have pioneered the automatic acquisition of lexical and semantic resources (such as verb classes). Some approaches are based on Harris's (1951) distribution hypothesis: syntactic structures with high occurrences can be used for identifying word clusters with common contexts (Lin and Pantel, 2001). Some others perform analysis from semantic networks (Green et al., 2004). Poibeau and Dutoit (2002) showed that both can be used in a complementary way.

Currently, our approach follows the first trend. Based on Hernandez and Grau (2003; 2004)'s proposal, we aim at explicitly using corpus characteristics such as its genre and domain features to reduce the quantity of considered data. In this paper we have explored various statistical measures which could be used as a filter for improving results obtained by the previous mentioned works.

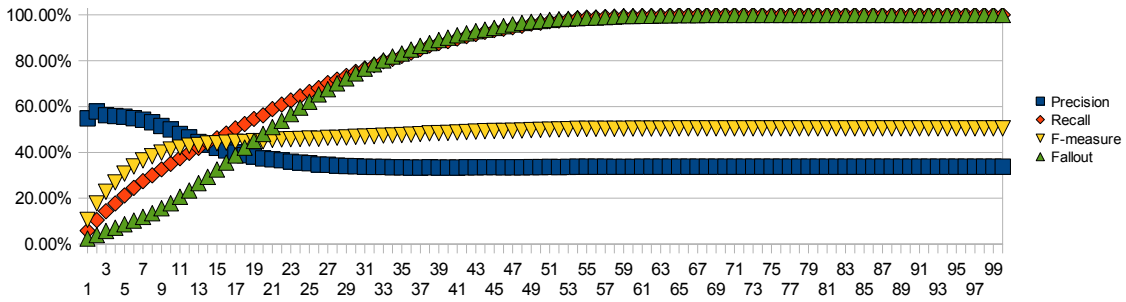


Figure 5: Tf.idf statistics

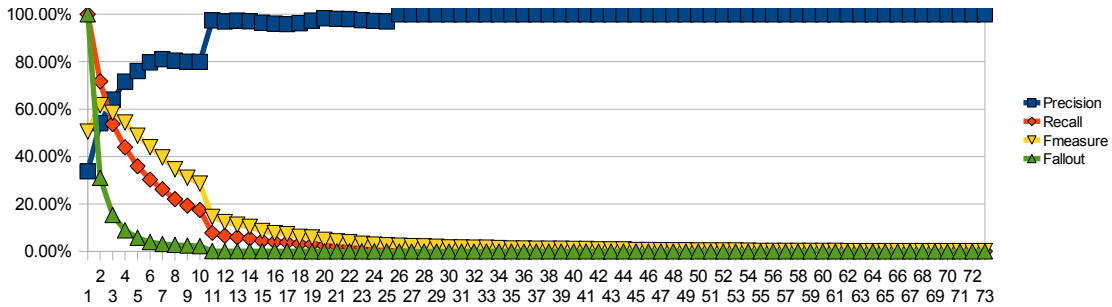


Figure 6: Inter-document frequency statistics

7 Conclusions and Future Work

In this paper we have presented a statistical approach for the extraction of a lexicon which contains the verbs and nouns that can be considered as candidates for use as predicates for the induction of predicate/argument structures that we call messages. Actually, the research presented here can be considered as the first step in a two-stages approach. The next step involves the use of clustering algorithms on the extracted lexicon which will provide the final clusters that will contain the predicates and arguments for the messages. This process is itself part of a larger process for the induction of predicate/argument structures. Apart from messages, such structures could as well be the structures that are associated with frame semantics, that is the frames and their associated semantic roles. Despite the great resemblances that messages and frames have, one of their great differences is the fact that messages were firstly introduced in the context of automatic multi-document summarization. By consequence they are meant to capture the most important information in a domain. Frames and semantic roles on the other hand, do not have this restriction and thus are more general. Nonetheless, it is our hope that the current research could ultimately be useful for the induction of frame semantics. In fact it is in our plans for the immediate future work to apply the same procedure in FrameNet annotated data¹⁰ in order to extract a vocabulary of verbs

¹⁰See http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=84

and nouns which will be characteristic of the different Linguistic Units (LUs) for the frames of FrameNet.

The proposed statistical measures are meant to be a first step towards a fully automated process of message induction. The immediate next step in the process involves the application of various unsupervised clustering techniques on the obtained lexicon in order to create the 48 different classes each one of which will represent a distinct vocabulary for the 48 different message types. We are currently experimenting with several algorithms such *K-means*, *Expectation-Minimization (EM)*, *Cobweb* and *Farthest First*. In addition to those clustering algorithms, we are also examining the use of various lexical association measures such as *Mutual Information*, *Dice coefficient*, χ^2 , etc. Although this approach will provide us with clusters of predicates and candidate arguments, still the problem of linking the predicates with their arguments remains. Undoubtedly, the use of more linguistically oriented techniques, such as syntactic analysis, is inevitable. We are currently experimenting with the use of a shallow parser (chunker) in order to identify the chunks that behave similarly in respect to a given cluster of predicates.

Concerning the evaluation of our approach, the highest F-measure score (61,87%) was given by the Collection Frequency statistical measure with a threshold value of 46%. This high F-measure though came at the cost of a high Fallout score (31.16%). Since the extracted lexicon will later be used as an input to a clustering algorithm, we would like to minimize as much as

possible the false positives. By consequence we have opted in using the Inter-document Frequency measure which presents an F-measure of 54.40% and a much more limited Fallout of 8.86%.

Acknowledgments

The authors would like to thank Konstantina Liontou and Maria Salapata for their help on the annotation of the messages, as well as the anonymous reviewers for their insightful and constructive comments.

References

- Stergos D. Afantenos, Irene Doura, Eleni Kapelou, and Vangelis Karkaletsis. 2004. Exploiting cross-document relations for multi-document evolving summarization. In G. A. Vouros and T. Panayiotopoulos, editors, *Methods and Applications of Artificial Intelligence: Third Hellenic Conference on AI, SETN 2004*, volume 3025 of *Lecture Notes in Computer Science*, pages 410–419, Samos, Greece, May. Springer-Verlag Heidelberg.
- Stergos D. Afantenos, Konstantina Liontou, Maria Salapata, and Vangelis Karkaletsis. 2005. An introduction to the summarization of evolving events: Linear and non-linear evolution. In Bernadette Sharp, editor, *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science, NLUCS 2005*, pages 91–99, Miami, Florida, USA, May. INSTICC Press.
- Stergos D. Afantenos, Vangelis Karkaletsis, Panagiotis Stamatopoulos, and Constantin Halatsis. 2008. Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. *Journal of Intelligent Information Systems*, 30(3):183–226, June.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.
- Fabio Ciravegna. 2001. Adaptive information extraction from text by rule induction and generalisation. In *17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 1251–1256, Seattle, USA, August.
- C. J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20–32.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- Rebecca Green, Bonnie J. Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from wordnet and Idoce. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 375–382, Barcelona, Spain, July.
- Zelig Harris. 1951. *Structural Linguistics*. University of Chicago Press.
- Nicolas Hernandez and Brigitte Grau. 2003. Automatic extraction of meta-descriptors for text description. In *International Conference on Recent Advances In Natural Language Processing (RANLP)*, Borovets, Bulgaria, 10-12 September.
- Nicolas Hernandez. 2004. *Détection et Description Automatique de Structures de Texte*. Ph.D. thesis, Université de Paris-Sud XI.
- Dekang Lin and Patrick Pantel. 2001. Induction of semantic classes from natural language text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 317–322, San Francisco, CA.
- Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May. Association for Computational Linguistics.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Thierry Poibeau and Dominique Dutoit. 2002. Inferring knowledge from a large semantic network. In *Proceeding of the Semantic networks workshop, during the Computational Linguistics Conference (COLING 2002)*, Taipei, Taiwan.
- Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Schefczyk. 2006. Framenet ii: Extended theory and practice. Unpublished manuscript; accessible at <http://framenet.icsi.berkeley.edu>.
- Horacio Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with sumum. *Computational Linguistics*, 28(4):497–526.
- Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.
- Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, June.
- Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.
- Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255, June.

Another look at indirect negative evidence

Alexander Clark

Department of Computer Science
Royal Holloway, University of London
alexcl@cs.rhul.ac.uk

Shalom Lappin

Department of Philosophy
King's College, London
shalom.lappin@kcl.ac.uk

Abstract

Indirect negative evidence is clearly an important way for learners to constrain over-generalisation, and yet a good learning theoretic analysis has yet to be provided for this, whether in a PAC or a probabilistic identification in the limit framework. In this paper we suggest a theoretical analysis of indirect negative evidence that allows the presence of ungrammatical strings in the input and also accounts for the relationship between grammaticality/acceptability and probability. Given independently justified assumptions about lower bounds on the probabilities of grammatical strings, we establish that a limited number of membership queries of some strings can be probabilistically simulated.

1 Introduction

First language acquisition has been studied for a long time from a theoretical point of view, (Gold, 1967; Niyogi and Berwick, 2000), but a consensus has not emerged as to the most appropriate model for learnability. The two main competing candidates, Gold-style identification in the limit and PAC-learning both have significant flaws.

For most NLP researchers, these issues are simply not problems: for all empirical purposes, one is interested in modelling the distribution of examples or the conditional distribution of labels given examples and the obvious solution – an $\epsilon - \delta$ bound on some suitable loss function such as the Kullback-Leibler Divergence – is sufficient (Horning, 1969; Angluin, 1988a). There may be some complexity issues involved with computing these approximations, but there is no debate about the appropriateness of the learning paradigm.

However, such an approach is unappealing to linguists for a number of reasons: it fails to draw a distinction between grammatical and ungrammatical sentences, and for many linguists the key

data are not the “performance” data but rather the “voice of competence” as expressed in grammaticality and acceptability judgments. Many of the most interesting sentences for syntacticians are comparatively rare and unusual and may occur with negligible frequency in the data.

We do not want to get into this debate here: in this paper, we will assume that there is a categorical distinction between grammatical and ungrammatical sentences. See (Schütze, 1996) for extensive discussion.

Within this view learnability is technically quite difficult to formalise in a realistic way. Children clearly are provided with examples of the language – so-called positive data – but the status of examples not in the language – negative data – is one of the endless and rather circular debates in the language acquisition literature (Marcus, 1993). Here we do not look at the role of corrections and other forms of negative data but we focus on what has been called indirect negative evidence (INE). INE is the non-occurrence of data in the primary linguistic data; informally, if the child does not hear certain ungrammatical sentences, then by their absence the child can infer that those strings are ungrammatical.

Indirect negative evidence has long been recognised as an important source of information (Pinker, 1979). However it has been surprisingly difficult to find an explicit learning theoretic account of INE. Indeed, in both the PAC and IIL paradigms it can be shown, that under the standard assumptions, INE cannot help the learner. Thus in many of these models, there is a sharp and implausible distinction between learning paradigms where the learner is provided systematically with every negative example, and those where the learner is denied any negative evidence at all. Neither of these is very realistic.

In this paper, we suggest a resolution for this conflict, by re-examining the standard learnability assumptions. We make three uncontroversial ob-

servations: first that the examples the child is provided with are *unlabelled*, secondly that there are a small proportion of ungrammatical sentences in the input to the child, and thirdly that in spite of this, the child does in fact learn.

We then draw a careful distinction between probability and grammaticality and propose a restriction on the class of distributions allowed to take account of the fact that children *are* exposed to some ungrammatical utterances. We call this the Disjoint Distribution Assumption: the assumption that the classes of distributions for different languages must be disjoint. Based on this assumption, we argue that the learner can infer lower bounds on the probabilities of grammatical strings, and that using these lower bounds allow a probabilistic approximation to membership queries of some strings.

On this basis we conclude that the learner does have some *limited* access to indirect negative evidence, and we discuss some of the limitations on this data and the implications for learnability.

2 Background

The most linguistically influential learnability paradigm is undoubtedly that of Gold (Gold, 1967). In this paradigm the learner is required to converge to exactly the right answer after a finite time. In one variant of the paradigm the learner is provided with only positive examples, and must learn on every presentation of the language. Under this paradigm no suprafinitive class of languages is learnable. If alternatively the learner is provided with a presentation of labelled examples, then pretty much anything is learnable, but clearly this paradigm has little relevance to the course of language acquisition.

The major problem with the Gold positive data paradigm is that the learner is required to learn under every presentation; given the minimal constraints on what counts as a presentation, this results in a model which is unrealistically hard. In particular, it is difficult for the learner to recover from an overly general hypothesis; since it is has only positive examples, such a hypothesis will never be directly contradicted.

Indirect negative evidence is the claim that the absence of sentences in the PLD can allow a learner to infer that those sentences are ungrammatical. As (Chomsky, 1981, p. 9) says:

A not unreasonable acquisition sys-

tem can be devised with the operative principle that if certain structures or rules fail to be exemplified in relatively simple expressions, where they would expect to be found, then a (possibly marked) option is selected excluding them in the grammar, so that a kind of “negative evidence” can be available even without corrections, adverse reactions etc.

While this informal argument has been widely accepted, and is often appealed to, it has so far not been incorporated explicitly into a formal model of learnability. Thus there are no learning models that we are aware of where positive learning results have been achieved using indirect negative evidence. Instead positive learnability results have typically used general probabilistic models of convergence without explicitly modelling grammaticality.

In what follows we will use the following notation. Σ is a finite alphabet, and Σ^* is the set of all finite strings over Σ . A (formal) language L is a subset of Σ^* . A distribution D over Σ^* is a function p_D from Σ^* to $[0, 1]$ such that $\sum_{w \in \Sigma^*} p_D(w) = 1$. We will write $\mathcal{D}(\Sigma^*)$ for the set of all distributions over Σ^* . The support of a distribution D is the set of strings with positive probability $supp(D) = \{w | p_D(w) > 0\}$.

3 Probabilistic learning

The solution is to recognise the probabilistic nature of how the samples are generated. We can assume they are generated by some stochastic process. On its own this says nothing – anything can be modelled by a stochastic process. To get learnability we will need to add some constraints.

Suppose the child has seen thousands of times sentences of the type “I am AP”, and “He is AP” where AP is an adjective phrase, but he has never heard anybody say “He am AP”. Intuitively it seems reasonable in this case to assume that the child can infer from this that sentences of the form “He am AP” are ungrammatical. Now, in the case of the Gold paradigm, the child can make no such inference. No matter how many millions or trillions of times he has heard other examples, the Gold paradigm does not allow any inference to be made from frequency. The teacher, or environment, is an adversary who might be deliberately withholding this data in order to confuse the

learner. The learner has to ignore this information.

However, in a more plausible learning environment, the learner can reason as follows. First, the number of times that the learner has observed sentences of the form “He am AP” is zero. From this, the learner can infer that sentences of this type are rare: i.e. that they are not very probable. Similarly from the high frequency of examples of the type “I am AP” and so on in the observed data, the learner can infer that the probability of these sentences is high.

The second step is that the learner can conclude from the difference in probability of these two similar sets of sentences, that there must be a difference in grammaticality between “He am AP” and “He is AP”, and thus that sentences of the type “He am AP” are ungrammatical.

It is important to recognise that the inference proceeds in two steps:

1. the first is the inference from low frequency in the observed data to low probability and
2. the second is the inference from *comparatively* low probability to ungrammaticality.

Both of these steps need justification, but if they are valid, then the learner can extract evidence about what is *not* in the language from stochastic evidence about what *is* in the language. The first step will be justified by some obvious and reasonable probabilistic assumptions about the presentation of the data; the second step is more subtle and requires some assumptions about the way the distribution of examples relates to the language being learned.

3.1 Stochastic assumptions

The basic assumption we make is that the samples are being generated randomly in some way; here we will make the standard assumption that each sentence is generated independently from the same fixed distribution, the Independently and Identically Distributed (IID) assumption. While this is a very standard assumption in statistics and probability, it has been criticised as a modelling assumption for language acquisition (Chater and Vitányi, 2007).

Here we are interested in the acquisition of syntax. We are therefore modelling the dependencies between words and phrases in sentences, but assuming that there are *no* dependencies between different sentences in discourse. That is to say, we

assume that the probability that a child hears a particular sentence does not depend on the previously occurring sentence. Clearly, there are dependencies between sentences. After questions, come answers; a polar interrogative is likely to be followed by a “yes” or a “no”; topics relate consecutive sentences semantically, and numerous other factors cause inter-sentential relationships and regularities of various types. Moreover, acceptability does depend a great deal on the immediate context. “Where did who go?” is marginal in most contexts; following “Where did he go?” it is perfectly acceptable. Additionally, since there are multiple people generating Child Directed Speech (CDS), this also introduces dependencies: each person speaks in a slightly different way; while a relative is visiting, there will be a higher probability of certain utterances, and so on. These correspond to a violation of the “identically” part of the IID assumption: the distribution will change in time.

The question is whether it is legitimate to neglect these issues in order to get some mathematical insight: do these idealising assumptions critically affect learnability? All of the computational work that we are aware of makes these assumptions, whether in a nativist paradigm, (Niyogi and Berwick, 2000; Sakas and Fodor, 2001; Yang, 2002) or an empiricist one (Clark and Thollard, 2004). We do need to make *some* assumptions, otherwise even learning the class of observed natural languages would be too hard. The minimal assumptions if we wish to allow any learnability under stochastic presentation are that the process generating the data is stationary and mixing. All we need is for the law of large numbers to hold, and for there to be rapid convergence of the observed frequency to the expectation. We can get this easily with the IID assumption, or with a bit more work using ergodic theory. Thus in what follows we will make the IID assumption; effectively using it as a place-holder for some more realistic assumption, based on ergodic processes. See for example (Gamarnik, 2003) for an extension of PAC analysis in this direction. The inference from low frequency to low probability follows from the minimal assumptions, specifically the IID, which we are making here.

4 Probability and Grammaticality

We now look at the second step in the probabilistic inference: how can the child go from low probabil-

ity to ungrammaticality? More generally the question is what is the relation between probability and grammaticality. There are lots of factors that affect probability other than grammaticality: length of utterance, lexical frequency, semantic factors and real world factors all can have an impact on probability.

Low probability on its own cannot imply ungrammaticality: if there are infinitely many grammatical sentences then there cannot be a lower bound on the probability: if all grammatical sentences have probability at least ϵ then there could be at most $1/\epsilon$ grammatical sentences which would make the language finite. A very long grammatical sentence can have very low probability, lower than a short ungrammatical sentence, so a less naive approach is necessary: the key point is that the probability must be *comparatively* low.

Since we are learning from unlabelled data, the only information that the child has comes from from the distribution of examples, and so the distribution must pick out the language precisely. To see this more clearly, suppose that the learner had access to an “Oracle” that would tell it the true probability of any string, and has no limit on how many strings it sees. A learner in this unrealistic model is clearly more powerful than any learner that just looks at a finite sample of the data. If this learner could not learn, then no real learner could learn on the basis of finite data.

More precisely for any language L we will have a corresponding set of distributions $\mathcal{D}(L)$, and we require the learner to learn under any of these distributions. What we require is that if we have two distinct languages L and L' then the two sets of distributions $\mathcal{D}(L)$ and $\mathcal{D}(L')$ must be disjoint, i.e. have no elements in common. If they did have a distribution D in common, then no learner could tell the two languages apart as the information being provided would be identical. Of course, given two distinct languages L and L' , it is possible that they intersect, that is to say that there are strings w in $L \cap L'$; a natural language example would be two related dialects of the same language such as some dialect of British English and some dialect of American; though the languages are distinct in formal terms, they are not disjoint, as there are sentences that are grammatical in both. When we consider the sets of distributions that are allowed for each language $\mathcal{D}(L)$ and $\mathcal{D}(L')$, we may find that there are elements $D \in \mathcal{D}(L)$ and $D' \in \mathcal{D}(L')$,

whose supports overlap, or even whose supports are identical, $\text{supp}(D) = \text{supp}(D')$, and we may well find that there are even some strings whose probabilities are identical; i.e. there may be a string w such that $p_D(w) = p_{D'}(w) > 0$. But what we do not allow is that we have a distribution D that is an element of both $\mathcal{D}(L)$ and $\mathcal{D}(L')$. If there were such an element, then when the learner was provided with samples drawn from this distribution, since the samples are unlabelled, there is absolutely no way that the learner could work out whether the target was L or L' ; the distribution would not determine the language. Therefore there must be a function from distributions to languages. We cannot have a single distribution that could be from two different languages. Let’s call this the disjoint distribution assumption (DDA): the assumption that the sets of distributions for distinct languages are disjoint.

Definition 1 *The Disjoint Distribution Assumption: If $L \neq L'$ then $\mathcal{D}(L) \cap \mathcal{D}(L') = \emptyset$.*

This assumption seems uncontroversial; indeed every proposal for a formal probabilistic model of language acquisition that we are aware of makes this assumption implicitly.

Now consider the convergence criterion: we wish to measure the error with respect to the distribution. There are two error terms, corresponding to false positives and false negatives. Suppose our target language is T and our hypothesis is H . Define $P_D(S)$ for some set S to be $\sum_{w \in S} p_D(w)$.

$$e^+ = P_D(H \setminus T) \quad (1)$$

$$e^- = P_D(T \setminus H) \quad (2)$$

We will require both of these error terms to converge to zero rapidly, and uniformly, as the amount of data the learner has increases.

5 Modelling the DDA

If we accept this assumption, then we will require some constraints on the sets of distributions. There are a number of ways to model this: the most basic way is to assume that strings have probability greater than zero if and only if the string is in the language. Formally, for all D in $\mathcal{D}(L)$

$$p_D(w) > 0 \Leftrightarrow w \in L \quad (3)$$

Here we clearly have a function from distributions to languages: we just take the support of the

distribution to be the language: for all D in $\mathcal{D}(L)$, $\text{supp}(D) = L$. Under this assumption alone however, indirect negative evidence will not be available.

That is because, in this situation, low probability does not imply ungrammaticality: only zero probability implies ungrammaticality. The fact that we have never seen a sentence in a finite sample of size n means that we can say that it is likely to have probability less than about $1/n$, but we cannot say that its probability is likely to be zero. Thus we can never conclude that a sentence is ungrammatical, if we make the assumption in Equation 3, and assume that there are no other limitations on the set of distributions. Since we have to learn for any distribution, we must learn even when the distribution is being picked adversarially. Suppose we have never seen an occurrence of a string; this could be because the probability has been artificially lowered to some infinitesimal quantity by the adversary to mislead us. Thus we gain nothing. Since there is no non-trivial lower bound on the probability of grammatical strings, effectively there is no difference between the requirement $p_D(w) > 0 \Leftrightarrow w \in L$ and the weaker condition $p_D(w) > 0 \Rightarrow w \in L$.

But this is not the only possibility: indeed, it is not a very good model at all. First, the assumption that ungrammatical strings have zero probability is false. Ungrammatical sentences, that is strings w , such that $w \notin L$, do occur in the environment, albeit with low probability. There are performance errors, poetry and songs, other children with less than adult competence, foreigners and many other potential sources of ungrammatical sentences. The orthodox view is that CDS is “unswervingly well-formed” (Newport et al., 1977): this is a slight exaggeration as a quick look at CHILDES (MacWhinney, 2000) will confirm. However, if we allow probabilities to be non-zero for ungrammatical sentences, and put no other restrictions on the distributions then the learner will fail on everything, since any distribution could be for any language.

Secondly, the convergence criterion becomes vacuous. As the probability of ungrammatical sentences is now zero, this means that $P_D(H \setminus T) = e^+ = 0$, and thus the vacuous learner that always returns the hypothesis Σ^* will have zero error. The normal way of dealing with this (Shvaytser, 1990) is to require the learner to hypothesize a subset of

the target. This is extremely undesirable, as it fails to account for the presence of over-generalisation errors in the child – or any form of production of ungrammatical sentences. On the basis of these arguments, we can see that this naive approach is clearly inadequate.

There are a number of other arguments why distribution free approaches are inappropriate here, even though they are desirable in standard applications of statistical estimation (Collins, 2005). First, the distribution of examples causally depends on the people who are uttering the examples who are native speakers of the language the learner is learning and use that knowledge to construct utterances. Second, suppose that we are trying to learn a class of languages that includes some infinite regular language L_r . For concreteness suppose it consists of $\{a^*b^*c^*\}$; any number of a’s followed by any number of b’s followed by any number of c’s. The learner must learn under any distribution: in particular it will have to learn under the distribution where every string except an infinitesimally small amount has the number of ‘a’s equal to the number of ‘b’s, or under the distribution where the number of occurrences of all three letters must be equal, or any other arbitrary subset of the target language. The adversary can distort the probabilities so that with probability close to one, at a fixed finite time, the learner will only see strings from this subset. In effect the learner has to learn these arbitrary subsets, which could be of much greater complexity than the language.

Indeed researchers doing computational or mathematical modelling of language acquisition often find it convenient to restrict the distributions in some way. For example (Niyogi and Berwick, 2000), in some computational modelling of a parameter-setting model of language acquisition say

In the earlier section we assumed that the data was uniformly distributed. ... In particular we can choose a distribution which will make the convergence time as large as we want. Thus the distribution-free convergence time for the three parameter system is infinite.

However, finding an alternative is not easy. There are no completely satisfactory ways of restricting the class of distributions, while maintaining the property that the support of the distribu-

tion is equal to the language. (Clark and Thollard, 2004) argue for limiting the class of distributions to those defined by the probabilistic variants of the standard Chomsky representations. While this is sufficient to achieve some interesting learning results, the class of distributions seems too small, and is primarily motivated by the requirements of the learning algorithm, rather than an analysis of the learning situation.

5.1 Other bounds

Rather than making the simplistic assumption that the support of the distribution must equal the language, we can instead make the more realistic assumption that every sentence, grammatical or ungrammatical, can in principle appear in the input and have non zero probability. In this case then we do not need to require the learner to produce a hypothesis that is a subset of the target, because if the learner overgeneralises, e^+ will be non-zero.

However, we clearly need to add some constraints to enforce the DDA. We can model this as a function from distributions to languages. It is obvious that grammaticality is correlated with probability in the sense that grammatical sentences are, broadly speaking, more likely than ungrammatical sentences; a natural way of articulating this is to say that there must be a real valued threshold function $g_D(w)$ such that if $p_D(w) > g_D(w)$ then $w \in L$. Using this we define the set of allowable distributions for a language L to be:

$$\mathcal{D}(L, g) = \{D : p_D(w) > g_D(w) \Leftrightarrow w \in L\} \quad (4)$$

Clearly this will satisfy the DDA. On its own this is vacuous – we have just changed notation, but this notation gives us a framework in which to compare some alternatives.

The original assumption that the support is equal to the languages in this framework then just has the simple form $g_D(w) = 0$. The naive constant bound we rejected above would be to have this threshold as a constant that depends neither on D nor on w i.e. for all w , $g_D(w) = \epsilon > 0$. Both of these bounds are clearly false, in the sense that they do not hold for natural distributions: the first because there are ungrammatical sentences with non-zero probability; the second because there are grammatical sentences with arbitrarily low probability. But the bound here need not be a constant, and indeed it can depend both on the distribution D and the word w .

5.2 Functional bound

We now look at variants of these bounds that provide a more accurate picture of the set of distributions that the child is exposed to. Recall that what we are trying to do is to characterise a range of distributions that is large enough to include those that the child will be exposed to. A slightly more nuanced way would be to have this as a very simple function of w , that ignores D , and is just a function of length. For example, we could have a simple uniform exponential model:

$$g_D(w) = \alpha_g \beta_g^{|w|} \quad (5)$$

This is in some sense an application of Harris’s idea of equiprobability (Harris, 1991):

whatever else there is to be said about the form of language, a fundamental task is to state the departures from equiprobability in sound- and word-sequences

Using this model, we do not assume that the learner is provided with information about the threshold g ; rather the learner will have certain, presumably domain general mechanisms that cause it to discard anomalies, and pay attention to significant deviations from equiprobability. We can view the threshold g as defining a bound on equiprobability; the role of syntax is to characterise these deviations from the assumption that all sequences are in some sense equally likely.

A more realistic model would depend also on D ; for example once could define these thresholds to depend on some simple observable properties of the distribution that could take account of lexical probabilities: more sophisticated versions of this bound could be derived from a unigram model, or a class-based model (Pereira, 2000).

Alternatively we could take account of the prefix and suffix probability of a string: for example, where for some $\alpha < 1$:¹

$$g_D(w) = \alpha \max_{uv=w} p_D(u\Sigma^*) p_D(\Sigma^*v) \quad (6)$$

6 Using the lower bound

Putting aside the specific proposal for the lower bound g , and going back to the issue of indirect

¹A prefix is just an initial segment of a string and has no linguistic and similarly for a suffix as the final segment.

negative evidence, we can see that the bound g is the missing piece in the inference: if we observe that a string w has zero frequency in our data set, then we can conclude it has low probability, say p ; if p is less than $g(w)$, then the string will be ungrammatical; therefore the inference from low probability to ungrammaticality in this case will be justified.

The bound here is justified independently: given the indubitable fact that there is a non-zero probability of ungrammatical strings in the child's input, and the DDA, which again seems unassailable, together with the fact that learners do learn some languages, it is a logical necessity that there is such a bound. This bound then justifies indirect negative evidence.

It is important to realise how limited this negative evidence is: it does not give the learner unlimited access to negative examples. The learner can only find out about sentences that would be frequent if they were grammatical; this may be enough to constrain overgeneralisation.

The most straightforward way of formalising this indirect negative evidence is with *membership queries* (Valiant, 1984; Angluin, 1988b). Membership queries are a model of learning where the learner, rather than merely passively receiving examples, can query an oracle about whether an example is in the language or not. In the model we propose, the learner can approximate a membership query with high probability by seeing the frequency of an example with a high g in a large sample. If the frequency is low, often zero, in this sample, then with high probability this example will be ungrammatical.

In particular given a functional bound, and some polynomial thresholds on the probability, and using Chernoff bounds we can simulate a polynomial number of membership queries, using large samples of data. Note that membership queries were part of the original PAC model (Valiant, 1984). Thus we can precisely define a limited form of indirect negative evidence.

In particular given a bound g , we can test to see whether a polynomial number of strings are ungrammatical by taking a large sample and examining their frequency.

The exact details here depend on the form of $g_D(w)$; if the bound depends on D in some respect the learner will need to estimate some aspect of D to compute the bound. This corresponds to

working out how probable the sentence would be if it were grammatical. In the cases we have considered here, given sufficient data, we can estimate $g_D(w)$ with high probability to an accuracy of ϵ_1 ; call the estimate $\hat{g}_D(w)$. We can also estimate the actual probability of the string with high probability again with accuracy ϵ_2 : let us denote this estimate by $\hat{p}_D(w)$. If $\hat{p}_D(w) + \epsilon_2 < \hat{g}_D(w) - \epsilon_1$, then we can conclude that $p_D(w) < g_D(w)$ and therefore that the sentence is ungrammatical. Conversely, the fact that a string has been observed once does not necessarily mean that it is grammatical. It only means that the probability is non-zero. For the learner to conclude that it is grammatical, s/he needs to have seen it enough times to conclude that the probability is above threshold. This will be if $\hat{p}_D(w) - \epsilon_2 > \hat{g}_D(w) + \epsilon_1$

Note that this may be slightly too weak and we might want to have a separate lower bound for grammaticality and upper bound for ungrammaticality. Otherwise if the distribution is such that many strings are very close to the boundary it will not be possible for the learner to determine whether they are grammatical or not.

We can thus define learnability with respect to a bound g that defines a set of distributions $\mathcal{D}(L, G)$. Thus this model differs from the PAC model in two respects: first the data is unlabelled, and secondly is is not distribution free.

Definition An algorithm A learns the class of languages \mathcal{L} if there is a polynomial p such that for every language $L \in \mathcal{L}$, where n is the size of the smallest representation of L , for all distributions $D \in \mathcal{D}(L, g)$ for all $\epsilon, \delta > 0$, when the algorithm A is provided with at least $p(n, \epsilon^{-1}, \delta^{-1}, \Sigma)$ unlabelled examples drawn IID from D , it produces with probability at least $1 - \delta$ a hypothesis H such that the error $P_D(H \setminus T \cup T \setminus H) < \epsilon$ and furthermore it runs in time polynomial in the total size of the sample.

7 Discussion

The unrealistic assumptions of the Gold paradigm were realised quite early on (Horning, 1969). It is possible to modify the Gold paradigm by incorporating a probabilistic presentation in the data and requiring the learner to learn with probability one. Perhaps surprisingly this does not change anything, if we put no constraints on the target distribution (Angluin, 1988a).

In particular given a presentation on which the

normal non-probabilistic learner fails, we can construct a distribution on which the probabilistic learner will fail. Thus allowing an adversary to pick the distribution is just as bad as allowing an adversary to pick the presentation. However, the distribution free assumption with unlabelled data cannot account for the real variety of distributions of CDS. In this model we propose restrictions on the class of distributions, motivated by the occurrence of ungrammatical sentences. This also means that we do not require a separate bound for over-generalisation. As a result, we conclude that there are limited amounts of negative evidence, and suggest that these can be formalised as a limited number of membership queries, of strings that would occur infrequently if they were ungrammatical.

To be clear, we are not claiming that this is a direct model of how children learn languages: rather we hope to get some insight into the fundamental limitations of learning from unlabelled data by switching to a more nuanced model. Here we have not presented any positive results using this model, but we observe that distribution dependent results for learning regular languages and some context free languages could be naturally modified to learn in this framework. We hope that the recognition of the validity of indirect negative evidence will direct attention away from the supposed problems of controlling overgeneralisation and towards the real problems: the computational complexity of inferring complex models.

References

- D. Angluin. 1988a. Identifying languages from stochastic examples. Technical Report YALEU/DCS/RR-614, Yale University, Dept. of Computer Science, New Haven, CT.
- D. Angluin. 1988b. Queries and concept learning. *Machine Learning*, 2(4):319–342, April.
- N. Chater and P. Vitányi. 2007. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163.
- N. Chomsky. 1981. Lectures on Government and Binding.
- Alexander Clark and Franck Thollard. 2004. Partially distribution-free learning of regular languages from positive samples. In *Proceedings of COLING*, Geneva, Switzerland.
- M. Collins. 2005. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In Harry Bunt, John Carroll, and Giorgio Satta, editors, *New Developments In Parsing Technology*, chapter 2, pages 19–55. Springer.
- D. Gamarnik. 2003. Extension of the PAC framework to finite and countable Markov chains. *IEEE Transactions on Information Theory*, 49(1):338–345.
- E. M. Gold. 1967. Language identification in the limit. *Information and control*, 10(5):447 – 474.
- Z.S. Harris. 1991. *A Theory of Language and Information: A Mathematical Approach*. Clarendon Press.
- James Jay Horning. 1969. *A study of grammatical inference*. Ph.D. thesis, Computer Science Department, Stanford University.
- B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates Inc, US.
- G.F. Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46(1):53–85.
- E.L. Newport, H. Gleitman, and L.R. Gleitman. 1977. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In *Talking to children: Language input and acquisition*, pages 109–149. Cambridge University Press.
- Partha Niyogi and Robert C. Berwick. 2000. Formal models for learning in the principle and parameters framework. In Peter Broeder and Jaap Murre, editors, *Models of Language Acquisition*, pages 225–243. Oxford University Press.
- F. Pereira. 2000. Formal grammar and information theory: Together again? In *Philosophical Transactions of the Royal Society*, pages 1239-1253. Royal Society, London.
- Steven Pinker. 1979. Formal models of language learning. *Cognition*, 7:217–282.
- W. Sakas and J.D. Fodor. 2001. The structural triggers learner. In *Language Acquisition and Learnability*, pages 172–233. Cambridge University Press.
- Carson T. Schütze. 1996. *The Empirical Base of Linguistics*. University of Chicago Press.
- H. Shvaytser. 1990. A necessary condition for learning from positive examples. *Machine Learning*, 5(1):101–113.
- L. Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134 – 1142.
- C.D. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press, USA.

Categorizing Local Contexts as a Step in Grammatical Category Induction

Markus Dickinson
Indiana University
Bloomington, IN USA
md7@indiana.edu

Charles Jochim
Indiana University
Bloomington, IN USA
cajochim@indiana.edu

Abstract

Building on the use of local contexts, or frames, for human category acquisition, we explore the treatment of contexts as categories. This allows us to examine and evaluate the categorical properties that local unsupervised methods can distinguish and their relationship to corpus POS tags. From there, we use lexical information to combine contexts in a way which preserves the intended category, providing a platform for grammatical category induction.

1 Introduction and Motivation

In human category acquisition, the immediate local context of a word has proven to be a reliable indicator of its grammatical category, or part of speech (e.g., Mintz, 2002, 2003; Redington et al., 1998). Likewise, category induction techniques cluster word types together (e.g., Clark, 2003; Schütze, 1995), using similar information, i.e., distributions of local context information. These methods are successful and useful (e.g. Koo et al., 2008), but in both cases it is not always clear whether errors in lexical classification are due to a problem in the induction algorithm or in what contexts count as identifying the same category (cf. Dickinson, 2008). The question we ask, then, is: what role does the context *on its own* play in defining a grammatical category? Specifically, when do two contexts identify the same category?

Many category induction experiments start by trying to categorize words, and Parisien et al. (2008) categorize *word usages*, a combination of a word and its context. But to isolate the effect the context has on the word, we take the approach of categorizing contexts as a first step towards clustering words. By separating out contexts for word clustering, we can begin to speak of better dis-

ambiguation models as a foundation for induction. We aim in this paper to thoroughly investigate what category properties contexts can or cannot distinguish by themselves.

With this approach, we are able to more thoroughly examine the categories used for evaluation. Evaluation of induction methods is difficult, due to the variety of corpora and tagsets in existence (see discussion in Clark, 2003) and the variety of potential purposes for induced categories (e.g., Koo et al., 2008; Miller et al., 2004). Yet improving the evaluation of category induction is vital, as evaluation does not match up well with grammar induction evaluation (Headden III et al., 2008). For many evaluations, POS tags have been mapped to a smaller tagset (e.g., Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), but there have been few criteria for evaluating the quality of these mappings. By isolating contexts, we can investigate how each mapping affects the accuracy of a method and the lexicon.

Using corpus annotation also allows us to explore the relation between induced categories and computationally or theoretically-relevant categories (e.g., Elworthy, 1995). While human category acquisition results successfully divide a lexicon into categories, these categories are not necessarily ones which are appropriate for many computational purposes or match theoretical syntactic analysis. This work can also serve as a platform to help drive the design of new tagsets, or refinement of old ones, by outlining which types of categories are or are not applicable for category induction.

After discussing some preliminary issues in section 2, in section 3 we examine to what extent contexts by themselves can distinguish different category properties and how this affects evaluation. Namely, we propose that corpus tagsets should be clear about identifying syntactic/distributional properties and about how tagset mappings for evaluation should outline how much information

is lost by mapping. In section 4, in more preliminary work, we add lexical information to contexts, in order to merge them together and see which still identify the same category.

2 Preliminaries

2.1 Background

Research on language acquisition has addressed how humans learn categories of words, and we use this as a starting point. Mintz (2002) shows that local context, in the form of a *frame* of two words surrounding a target word, leads to the target’s categorization in adults, and Mintz (2003) shows that frequent frames supply category information in child language corpora. A frame is not decomposed into its left and right sides (cf., e.g., Redington et al., 1998; Clark, 2003; Schütze, 1995), but is taken as their joint occurrence (Mintz, 2003).¹

For category acquisition, *frequent frames* are used, those with a frequency above a certain threshold. These predict category membership, as the set of words appearing in a given frame should represent a single category. The frequent frame *you ... it*, for example, largely identifies verbs, as shown in (1), taken from child-directed speech in the CHILDES database (MacWhinney, 2000). For frequent frames in six subcorpora of CHILDES, Mintz (2003) obtains both high type and token accuracy in categorizing words.

- (1) a. you **put** it
b. you **see** it

The categories do not reflect fine-grained linguistic distinctions, though, nor do they fully account for ambiguous words. Indeed, accuracies slightly degrade when moving from “Standard Labeling”² to the more fine-grained “Expanded Labeling,”³ from .98 to .91 in token accuracy and from .93 to .91 in type accuracy. In scaling the method beyond child-directed speech, it would be beneficial to use annotated data, which allows for ambiguity and distinguishes a word’s category across corpus instances. Furthermore, even though many frames identify the same category,

¹This use of *frame* is different than that used for subcategorization frames, which are also used to induce word classes (e.g., Korhonen et al., 2003).

²Categories = noun, verb, adjective, preposition, adverb, determiner, wh-word, *not*, conjunction, and interjection.

³Nouns split into nouns and pronouns; verbs split into verbs, auxiliaries, and copula

the method does not thoroughly specify how to relate them.

It has been recognized for some time that wider contexts result in better induction models (e.g., Parisien et al., 2008; Redington et al., 1998), but many linguistic distinctions rely on lexical information that cannot be inferred from additional context (Dickinson, 2008), so focusing on short contexts can provide many insights. The use of frames allows for frequent recurrent contexts and a way to investigate corpus categories, or POS tags (cf., e.g., Dickinson and Jochim, 2008). An added benefit of starting with this method is that it can be converted to a model of online acquisition (Wang and Mintz, 2007). For this paper, however, we only investigate the type of information input into the model.

2.2 Some definitions

Frequency The core idea of using frames is that words used in the same context are associated with each other, and the more often these contexts occur, the more confidence we have that the frame indicates a category. Setting a threshold to obtain the 45 most frequent frames in each subcorpus (about 80,000 words on average), (Mintz, 2003) allows a frame to occur often enough to be meaningful and have a variety of target words in the frame.

To determine what category properties frames pinpoint (section 3), we use two thresholds to define *frequent*. Singly occurring frames cannot provide any information about groupings of words, so we first consider frames that occur more than once. This gives a large number of frames, covering much of the corpus (about 970,000 tokens), but frames with few instances have very little information. For the other threshold, frequent frames are those which have a frequency of 200, about 0.03% of the total number of frames in the corpus. One could explore more thresholds, but for comparing tagset mappings, these provide a good picture. The higher threshold is appropriate for combining contexts (section 4), as we need more information to tell whether two frames behave similarly.

Accuracy To evaluate, we need a measure of the accuracy of each frame. Mintz (2003) and Redington et al. (1998) calculate accuracy by counting all pairs of words (types or tokens) that are from the same category, divided by all possible pairs of words in a grouping. This captures the idea that each word should have the same category as every

other word in its category set.

Viewing the task as disambiguating contexts (see section 3), however, this measurement does not seem to adequately represent cases with a majority label. For example, if three words have the tag X and one Y , pairwise comparison results in an accuracy of 50%, even though X is dominant. To account for this, we measure the precision of the most frequent category instances among all instances, e.g., 75% for the above example (cf. the notion of *purity* in Manning et al., 2008). Additionally, we only use measurements of token precision. Token precision naturally handles ambiguous words and is easy to calculate in a POS-annotated corpus.

3 Categories in local contexts

In automatic category induction, a category is often treated as a set, or cluster, of words (Clark, 2003; Schütze, 1995), and category ambiguity is represented by the fact that words can appear in more than one set. Relatedly, one can cluster *word usages*, a combination of a word and its context (Parisien et al., 2008). An erroneous classification occurs when a word is in an incorrect set, and one source of error is when the contexts being treated as indicative of the same category are actually ambiguous. For example, in a bigram model, the context *be _* identifies nouns, adjectives, and verbs, among others.

Viewed in this way, it is important to gauge the precision of contexts for distinguishing a category (cf. also Dickinson, 2008). In other words, how often does the same context identify the same category? And how fine-grained is the category that the context distinguishes? To test whether a frame defines a single category in non-child-directed speech, we focus on which categorical properties frames define, and for this we use a POS-annotated corpus. Due to its popularity for unsupervised POS induction research (e.g., Goldberg et al., 2008; Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008) and its often-used tagset, for our initial research, we use the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1993), with 36 tags (plus 9 punctuation tags), and we use sections 00-18, leaving held-out data for future experiments.⁴

Defining frequent frames as those occurring at

⁴Even if we wanted child-directed speech, the CHILDES database (MacWhinney, 2000) uses coarse POS tags.

least 200 times, we find 79.5% token precision. Additionally, we have 99 frames, identifying 14 types of categories as the majority tag (common noun (NN) being the most prevalent (37 frames)). For a threshold of 2, we have 77.3% precision for 67,721 frames and 35 categories.⁵ With precision below 80%, we observe that frames are not fully able to disambiguate these corpus categories.

3.1 Frame-defined categories

These corpus categories, however, are composed of a variety of morphological and syntactic features, the exact nature of which varies from tagset to tagset. By merging different tags, we can factor out different types of morphological and syntactic properties to determine which ones are more or less easily identified by frames. Accuracy will of course improve by merging tags; what is important is for which mappings it improves.

We start with basic categories, akin to those in Mintz (2003). Despite the differences among tagsets, these basic categories are common, and merging POS tags into basic categories can show that differences in accuracy have more to do with stricter category labels than language type. We merged tags to create basic categories, as in table 1 (adapted from Hepple and van Genabith (2000); see appendix A for descriptions).⁶

Category	Corpus tags
Determiner	DT, PDT, PRP\$
Adjective	JJ, JJR, JJS
Noun	NN, NNS, PRP, NNP, NNPS
Adverb	RB, RBR, RBS
Verb	MD, VB, VBD, VBG, VBN, VBP, VBZ
<i>Wh</i> -Det.	WDT, WP\$

Table 1: Tag mappings into basic categories

These broader categories result in the accuracies in table 2, and we also record accuracies for the similar PTB-17 tagset used in a variety of unsupervised tagging experiments (Smith and Eisner, 2005), which mainly differs by treating VBG and VBN uniquely. With token precision around 90%, it seems that frame-based disambiguation is generally identifying basic categories, though with less

⁵LS (List item marker) is not identified; UH (interjection) appears in one repeating frame, and SYM (symbol) in two.

⁶The 13 other linguistic tags were not merged, i.e., CC, CD, EX, FW, IN, LS, POS, RP, SYM, TO, UH, WP, WRB.

accuracy than in Mintz (2003).

	≥ 2	≥ 200
Orig.	77.3%	79.5%
Merged	85.9%	91.0%
PTB-17	85.1%	89.7%

Table 2: Effect of mappings on precision

But which properties of the tagset do the frame contexts accurately capture and which do they not? To get at this question, we explore linguistically-motivated mappings between the original tagset and the fully-merged tagset in table 1. Given the predominance of verbs and nouns, we focus on distinguishing linguistic properties within these categories. For example, simply by merging nouns and leaving all other original tags unchanged, we move from 79.5% token precision to 88.4% (for the threshold of 200).

Leaving all other mappings as in table 1, we merge nouns and verbs along two dimensions: their common syntactic properties or their common morphological properties. Ideally, we prefer frames to pick out syntactic properties, since morphological properties can assumedly be determined from word-internal properties (see Clark, 2003; Christiansen and Monaghan, 2006).

Specifically, we can merge nouns by *noun type* (PRP [pronoun], NN/NNS [common noun], NNP/NNPS [proper noun]) or by *noun form*, in this case based on grammatical number (PRP [pronoun], NN/NNP [singular noun], NNS/NNPS [plural noun]). We can merge verbs by *finiteness* (MD [modal], VBP/VBZ/VBD [finite verb], VB/VBG/VBN [nonfinite verb]) or by *verb form* (MD [modal], VB/VBP [base], VBD/VBN [-ed], VBG [-ing], VBZ [-s]). In the latter case, verbs with consistently similar forms are grouped—e.g., *see* can be a baseform (VB) or a present tense verb (VBP).

The results are given in tables 3 and 4. We find that merging verbs by finiteness and nouns by noun type results in higher precision. This confirms that contexts can better distinguish syntactic, but not necessarily morphological, properties. As we will see in the next section, this mapping also maintains distinctions in the lexicon. Such use of local contexts, along with tag merging, can be used to evaluate tagsets which claim to be distributional (see, e.g., Dickinson and Jochim, 2008).

It should be noted that we have only explored

	Noun type	Noun form
Finiteness	82.9%	81.2%
Verb form	81.2%	79.5%

Table 3: Mapping precision (freq. ≥ 2)

	Noun type	Noun form
Finiteness	86.4%	85.3%
Verb form	84.5%	83.4%

Table 4: Mapping precision (freq. ≥ 200)

category mappings which merge tags, ignoring possible splits. While splitting a tag like TO (*to*) into prepositional and infinitival uses would be ideal, we do not have the information automatically available. We are thus limited in our evaluation by what the tagset offers. Some tag splits can be automatically recovered (e.g., splitting PRP based on properties such as person), but if it is automatically recoverable from the lexicon, we do not necessarily need context to identify it, an idea we turn to in the next section.

3.2 Evaluating tagset mappings

Some of the category distinctions made by frames are more or less important for the context to make. For example, it is detrimental if we conflate VB and VBP because this is a prominent ambiguity for many words (e.g., *see*). On the other hand, there are no words which can be both VBP (e.g., *see*) and VBZ (e.g., *sees*). Ideally, induction methods would be able to distinguish all these cases—just as they often make distinctions beyond what is in a tagset—but there are differences in how problematic the mappings are. If we group VB and VBP into one tag, there is no way to recover that distinction; for VBP and VBZ, there are at least different words which inherently take the different tags.

Thus, a mapping is preferred which does not conflate tags that vary for individual words. To calculate this, we compare the original lexicon with a mapped lexicon and count the number of words which lose a distinction. Consider the words *accept* and *accepts*: *accept* varies between VB and VBP; *accepts* is only VBZ. When we map tags based on verb form, we count 1 for *accept*, as VB and VBP are now one tag (Verb). When we map verbs based on finiteness, we count 0 for these two words, as *accept* still has two tags (V-nonfin, V-fin) and *accepts* has one tag (V-fin).

We evaluate our mappings in table 5 by enumerating the number of word types whose distinctions are lost by a particular mapping (out of 44,520 word types); we also repeat the token precision values for comparison. Perhaps unsurprisingly, grouping words based on form results in high confusability (cf. the discussion of *see* in section 3.1). On the other hand, merging nouns by type and verbs by finiteness results in something of a balance between precision and non-confusability. It is thus these types of categorizations which we can reasonably expect induction models to capture.

Mapping	Lost tags	Precision	
		≥ 2	≥ 200
All mappings	3003	85.9%	91.0%
PTB-17	2038	85.1%	89.7%
N. form/V. form	2699	79.5%	83.4%
N. type/V. form	2148	81.2%	84.5%
N. form/Finite	951	81.2%	85.3%
N. type/Finite	399	82.9%	86.4%
No mappings	0	77.3%	79.5%

Table 5: Confusable word types

For induction evaluation, in addition to an accuracy metric, a metric such as the one we have just proposed is important to gauge how much corpus annotation information is lost when performing tagset mappings. For example, the PTB-17 mapping (Smith and Eisner, 2005) is commonly used for evaluating category induction (Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), yet it loses distinctions for 2038 words.

We could also define mappings which lose no distinctions in the lexicon. Initial experiments show that this allows no merging of nouns, and that the resulting precision is only minimally better than no mapping at all. We should also note that the number of confusable words may be too high, given errors in the lexicon (cf. Dickinson, 2008). For example, removing tags occurring less than 10% of the time for a word results in only 305 confusable words for the Noun type/Finiteness (NF) mapping and 1575 for PTB-17.

4 Combining contexts

We have narrowly focused on identical contexts, or frames, for identifying categories, but this could leave us with as many categories as frames (67,721 for ≥ 2 , 99 for ≥ 200 , instead of 35 and 30). We need to reduce the number of categories without

inappropriately merging them (cf. the notion of “completeness” in Mintz, 2003; Christiansen and Monaghan, 2006). Thus far, we have not utilized a frame’s target words; we turn to these now, in order to better gauge the effectiveness of frames for identifying categories. Although the work is somewhat preliminary, our goal is to continue to investigate when contexts identify the same category. This merging of contexts is different than clustering words (e.g., Clark, 2000; Brown et al., 1992), but is applicable, as word clustering relies on knowing which contexts identify the same category.

4.1 Word-based combination

On their own, frames at best distinguish only very broad categorical properties. This is perhaps unsurprising, as the finer-grained distinctions in corpora seem to be based on lexical properties more than on additional context (see, e.g., Dickinson, 2008). If we want to combine contexts in a way which maps to corpus tagsets, then, we need to examine the target words. It is likely that two sets share the same tag if they contain the same words (cf. overlap in Mintz, 2003). In fact, the more a frame’s word set overlaps with another’s word set, the more likely it is unambiguous in the first place, as the other set provides corroborating evidence. Therefore, we use overlap of frames’ word sets as a criterion to combine them.

This allows us to combine frames which do not share context words. For example, in (2) we find frames identifying baseform verbs (VB) (2a) and frames identifying cardinal numbers (CD) (2b), despite having a variety of context words. Their target word sets, however, are sufficiently similar.

- (2) a. will __ to, will __ the, to __ the, to __ up,
would __ the, to __ their, n’t __ the,
to __ a, to __ its, to __ that, to __ to
- b. or __ cents, \$ __ million, rose __ %,
a __ %, about __ %, to __ %, \$ __ a,
\$ __ billion

By viewing frames as categories, in the future we could also investigate splitting categories, based on subsets of words, morphological/phonological cues (e.g., Christiansen and Monaghan, 2006), or on additional context words, better handling frames that are ambiguous.

Calculating overlap We merge frames whose word sets overlap, using a simple weighted fre-

quency distance metric. We define sufficient overlap as the case where a given percent of the words in one frame’s word set are found in the other’s word set. We define this test in either direction, as smaller sets can be a subset of a larger set. For example, the frames *the ... on* (224 tokens) and *the ... of* (4304 tokens) have an overlap of 78 tokens; overlap here is 34.8% (78/224). While we could use a more sophisticated form of clustering (see, e.g., Manning et al., 2008), this will help determine the viability of this general approach.

Of course, two sets may share a category with relatively few shared words, and so we transitively combine sets of contexts. If the overlap of frames *A* and *B* meet our overlap criterion and the overlap of frames *A* and *C* also meet the criterion, then all three sets are merged, even if *B* and *C* have only a small amount of overlap.⁷

Using the threshold of 200, we test criteria of 30%, 40%, and 50% overlap and consider the frames’ overlap calculated as a percentage of word types or as a percentage of word tokens. For example, if a word type occurs 10 times in one word set and 20 in the other, the overlap of types is 1, and the overlap of tokens is 10. Token overlap better captures similarities in distributions of words.

4.2 Evaluation

Table 6 shows the number of categories for the 30%, 40%, and 50% type-based (TyB) and token-based (ToB) overlap criteria for merging. As we can see, the overlap based on tokens in word sets results in more categories, i.e., fewer merges.

%	TyB	ToB
50%	59	75
40%	42	64
30%	27	50

Table 6: Number of categories by condition

The precision of each of these criteria is given in table 7, evaluating on both the original tagset and the noun type/finiteness (*NF*) mapping. We can see that the token-based overlap is consistently more accurate than type-based overlap, and there is virtually no drop in precision for any of the token-based conditions.⁸ Thus, for the rest of the evaluation, we use only the token-based overlap.

⁷We currently do not consider overlap of already merged sets, e.g., between *A+B* and *C*.

⁸Experiments at 20% show a noticeable drop in precision.

%	Tags	Frames	TyB	ToB
50%	Orig.	79.5%	76.4%	79.5%
	NF	86.4%	82.8%	86.4%
40%	Orig.	79.5%	75.7%	79.3%
	NF	86.4%	81.8%	86.1%
30%	Orig.	79.5%	74.7%	79.1%
	NF	86.4%	81.7%	86.1%

Table 7: Precision of merged frames

We mentioned that if frame word sets overlap, the less ambiguous their category should be. We check this by looking at the difference between merged and unmerged frames, as shown in table 8. The number of categories are also given in parentheses; for example, for 30% overlap, 41 frames are unmerged, and the remaining 58 make up 9 categories. These results confirm for this data that frames which are merged have a higher precision.

	Merged	Unmerged	Overall
50%	93.4% (7)	79.9% (68)	86.4% (75)
40%	89.7% (10)	81.1% (54)	86.1% (64)
30%	89.7% (9)	77.4% (41)	86.1% (50)

Table 8: Precision of merged & unmerged frames for NF mapping (with number of categories)

But are we only merging a select, small set of words? To gauge this, we measure how much of the corpus is categorized by the 99 most frequent frames. Namely, 46,874 tokens occur as targets in our threshold of 99 frequent frames out of 663,608 target tokens in the entire corpus,⁹ a recall of 7.1%. Table 9 shows some recall figures for the frequent frames. There are 9621 word types in the set of target words for the 99 frequent frames, which is 27.2% of the target lexicon. Crucially, though, these 9621 are realized as 523,662 target tokens in the corpus, or 78.9%. The words categorized by the frequent frames extend to a large portion of the corpus (cf. also Mintz, 2003).

	Tokens	Types	Coverage
Merged (30%)	5.0%	20.0%	61.5%
Unmerged (30%)	2.0%	11.5%	65.9%
Total Overlap	7.1%	27.2%	78.9%

Table 9: Recall of frames

⁹Because we remove frames which contain punctuation, the set of target tokens is a subset of all words in the corpus.

4.2.1 Qualitative analysis

To better analyze what is happening for future work, we look more closely at 30% overlap. Of the 58 frames merged into 9 categories, 54 of them have the same majority tag after merging. The four frames which get merged into a different category are worth investigating, to see the method’s limitations and potential for improvement.

Of the four frames which lose their majority tag after merging, two can be ignored when mapping to the NF tags. The frame *it ... the* with majority tag VBZ becomes VBD when merged, but both are V-fin. Likewise, *n’t ... to* changes from VB to VBN, both cases of V-nonfin. The third case reveals an evaluation problem with the original tagset: the frames *million ... \$* (IN) and *% ... \$* (TO) are merged into a category labeled TO. The tag TO is for the word *to* and is not split into prepositional and infinitival uses. Corpus categories such as these, which overlap in their definitions yet cannot be merged (due to their non-overlapping uses), are particularly problematic for evaluation.

The final case which does not properly merge is the most serious. The frame *is ... the* (37% of tokens as preposition (IN)) merges with *is ... a* (41% of tokens as VBG); the merged VBG category has an precision of 34%. The distribution of tags is relatively similar, the highest percentages being for IN and VBG in both. This highlights the point made earlier, that more information is needed, to split the word sets.

4.2.2 TIGER Corpus

To better evaluate frequent frames for determining categories, we also test them on the German TIGER corpus (Brants et al., 2002), version 2, to see how the method handles data with freer word order and more morphological complexity. We use the training data, with the data split as in Dubey (2004). The frequency threshold for the WSJ (0.03% of all frames) leaves us with only 60 frames in the TIGER corpus, and 51 of these frames have a majority tag of NN.¹⁰ Thus, we adjusted the threshold to 0.02% (102 minimum occurrences), thereby obtaining 119 frequent frames, with a precision of 82.0%. For the 30% token-based overlap (the best result for English), frames merged into 81 classes, with 79.1% precision. These precision figures are on a par with

¹⁰We use no tagset mappings for our TIGER experiments.

English (cf. table 7).¹¹ Part of this might be due to the fact that NN is still a large majority (76% of the frames). Additionally, we find that, although the frame tokens make up only 5.2% of the corpus and the types make up 15.9% of the target lexicon, those types correspond to 67.2% of the target corpus tokens.

5 Summary and Outlook

Building on the use of frames for human category acquisition, we have explored the benefits of treating contexts—in this case, frames—as categories and analyzed the consequences. This allowed us to examine a way to evaluate tagset mappings and provide feedback on distributional tagset design. From there, we explored using lexical information to combine contexts in a way which generally preserves the intended category.

We evaluated this on English and German, but, to fully verify our findings, a high priority is to perform similar experiments on more corpora, employing different tagsets, for different languages. Additionally, we need to expand the definition of a context to more accurately categorize contexts, while at the same time not lowering recall.

Acknowledgements

We wish to thank the Indiana University Computational Linguistics discussion group for feedback, as well as the three anonymous reviewers.

A Some Penn Treebank POS tags

DT	Determiner
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
PRP	Personal pronoun
PRP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP\$	Possessive wh-pronoun

¹¹Interestingly, thresholds of 20% and 10% result in similarly high precision.

References

- Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). The TIGER Treebank. In *Proceedings of TLT-02*. Sozopol, Bulgaria.
- Brown, Peter F., Peter V. deSouza, Robert L. Mercer, T. J. Watson, Vincent J. Della Pietra and Jenifer C. Lai (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4), 467–479.
- Christiansen, Morten H. and Padraic Monaghan (2006). Discovering verbs through multiple-cue integration. In *Action Meets Word: How Children Learn Verbs*, Oxford: OUP.
- Clark, Alexander (2000). Inducing Syntactic Categories by Context Distribution Clustering. In *Proceedings of CoNLL-00*. Lisbon, Portugal.
- Clark, Alexander (2003). Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of EACL-03*. Budapest.
- Dickinson, Markus (2008). Representations for category disambiguation. In *Proceedings of Coling 2008*. Manchester.
- Dickinson, Markus and Charles Jochim (2008). A Simple Method for Tagset Comparison. In *Proceedings of LREC 2008*. Marrakech, Morocco.
- Dubey, Amit (2004). Statistical Parsing for German: Modeling syntactic properties and annotation differences. Ph.D. thesis, Saarland University, Germany.
- Elworthy, David (1995). Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*. Dublin.
- Goldberg, Yoav, Meni Adler and Michael Elhadad (2008). EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start). In *Proceedings of ACL-08*. Columbus, OH.
- Goldwater, Sharon and Tom Griffiths (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL-07*. Prague.
- Headden III, William P., David McClosky and Eugene Charniak (2008). Evaluating Unsupervised Part-of-Speech Tagging for Grammar Induction. In *Proceedings of Coling 2008*. Manchester.
- Hepple, Mark and Josef van Genabith (2000). Experiments in Structure-Preserving Grammar Compaction. In *1st Meeting on Speech Technology Transfer*. Seville, Spain.
- Koo, Terry, Xavier Carreras and Michael Collins (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of ACL-08*. Columbus, OH.
- Korhonen, Anna, Yuval Krymolowski and Zvika Marx (2003). Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of ACL-03*. Sapporo.
- MacWhinney, Brian (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third edn.
- Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*. CUP.
- Marcus, M., Beatrice Santorini and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.
- Miller, Scott, Jethran Guinness and Alex Zamanian (2004). Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of HLT-NAACL 2004*. Boston, MA.
- Mintz, Toben H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* 30, 678–686.
- Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.
- Parisien, Christopher, Afsaneh Fazly and Suzanne Stevenson (2008). An Incremental Bayesian Model for Learning Syntactic Categories. In *Proceedings of CoNLL-08*. Manchester.
- Redington, Martin, Nick Chater and Steven Finch (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22(4), 425–469.
- Schütze, Hinrich (1995). Distributional Part-of-Speech Tagging. In *Proceedings of EACL-95*. Dublin, Ireland.
- Smith, Noah A. and Jason Eisner (2005). Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of ACL'05*. Ann Arbor, MI.
- Toutanova, Kristina and Mark Johnson (2008). A Bayesian LDA-based Model for Semi-Supervised Part-of-speech Tagging. In *Proceedings of NIPS 2008*. Vancouver.
- Wang, Hao and Toben H. Mintz (2007). A Dynamic Learning Model for Categorizing Words Using Frames. In *Proceedings of BUCLD 32*. pp. 525–536.

Darwinised Data-Oriented Parsing – Statistical NLP with added Sex and Death

Dave Cochran

Cognitive Systems Group,
School of Computer Science,
University of St. Andrews

davec@cs.st-andrews.ac.uk

Abstract

We present the Darwinised Data-Oriented Parsing algorithm, an incremental, dynamic form of Data-Oriented Parsing, in which exemplars are used as replicators, subject to a selection pressure towards generalisability.¹

1 Introduction

Data-Oriented Parsing (DOP) is a state-of-the-art approach to both supervised and unsupervised parsing (Bod 1992, 1998, 2006a, 2006b, 2007a, 2007b, Zollman and Sima'an 2005), which has mostly been developed within a technologically-oriented computer science context. Recent work has highlighted some interesting cognitive properties of the Data-Oriented approach (Borensztajn, Zuidema & Bod 2008, Bod 2008). However, these studies have mostly focused on the static properties of the DOP probability model. Here, we present the first attempt at a dynamic, incremental Data-Oriented model which can address the time course of language learning, rather than just the outcome; Darwinised DOP.

2 Data-Oriented Parsing

2.1 Supervised DOP

Data-Oriented Parsing (DOP) is a paradigm in Natural Language Processing in which linguistic knowledge is represented as fragmentable, re-combinable exemplars of concrete previous experience, usually in the form of *trees*. What crucially distinguishes DOP from other approaches is the fact that fragments of arbitrary size are

used, ranging, in the case of the usual tree-structures, from depth-1 context-free rewrite rules to entire trees, and all points in between; this gives it the power to pick up on whatever statistical patterns are present in the data, to a considerable extent bypassing of the researcher's theoretical prejudices. Moreover, it allows these regularities to be exploited *without being represented*. DOP was first proposed by Scha (1990), and implemented and developed by Bod (1992, 1998).

The simplest manifestation of DOP is DOP1, as described in Bod (1998 p12-23 and 40-50), though more sophisticated versions exist. The parser uses a large corpus of natural language strings annotated with labeled, ordered tree-structures, divided into a training corpus and a smaller corpus against which the parser is tested. The parser uses every possible subtree (of unlimited depth) of all the available trees, constrained only by the following wellformedness criteria (Fig. 1).

- Every subtree must be of at least depth 1.
- Sister relationships must be preserved: that is, either all or none of the daughters of a given node may be extracted, but not only some.

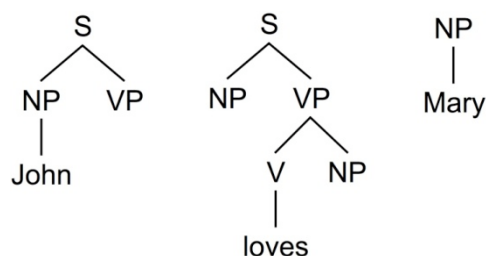


Figure 1: Well-formed subtrees of a parse of "John likes Mary"

¹ The author thanks Rens Bod, Mark-Jan Nederhof and three anonymous reviewers for exceedingly helpful comments, suggestions and references.

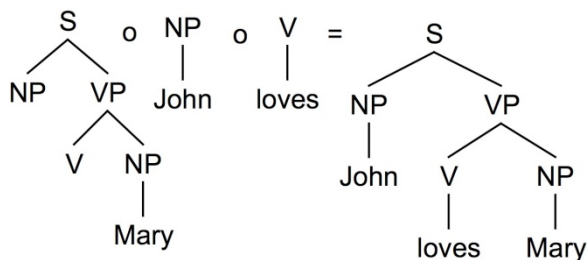


Fig 2: A sequence of substitutions comprising a derivation of “John likes Mary”. \circ is the operator for leftmost non-terminal leaf-node substitution.

The parser is given test corpus strings and builds up new parse-trees for these using the fragments available to it from the training corpus (Fig. 2), starting with a fragment with an S-node at the top, and then, for each nonterminal leaf-node, working rightwards, substituting in additional subtrees, the topmost node of which must carry the same label as the node to be substituted. (see figure 2.2).

In DOP research it is necessary to distinguish between *parses* and *derivations*. A parse is the tree structure expressed over a string; a derivation is the particular sequence of subtree substitutions by which it was constructed. When parsing with probabilistic context-free grammars (PCFG’s, see Manning and Schütze 1999, pp.381-405; note that a PCFG is equivalent to a DOP grammar in which subtree depth has been restricted to 1), there is a one-to-one mapping between parses and derivations, because all non-terminal nodes (nodes which have daughters in the completed parse) *must* be substitution sites. In DOP, subtrees can be of any depth, and so in any given derivation, any subset of the non-terminal nodes could have been substitution sites, while the remainder will not have been. As such, if a parse contains N many non-terminal nodes, it will have 2^N many derivations.

For each subtree t , its probability $P(t)$ is its total frequency $|t|$ of occurrence in the training corpus over the summed corpus frequency of subtrees with the same root node;²

$$P(t) = \frac{|t|}{\sum_{\{t':r(t)=r(t')\}}|t'|} \quad (1)$$

...where $r(t)$ and $r(t')$ are the node-labels on the root-nodes of subtrees t and t' .

² Note that although, beside the node-label on the substitution site, the input to be parsed is also a constraint on the selection of subtrees for substitutions, these constraints are not factored in to the calculation of probabilities. That is to say, the probability used is that of the subtree, not the substitution.

The probability of a derivation is the product of the probabilities of its subtrees (note that \circ is the notation for the substitution operation; thus $t_1 \circ \dots \circ t_n$ is the sequence of substitutions, which together comprise the derivation);

$$P(t_1 \circ \dots \circ t_n) = \prod_i P(t_i) \quad (2)$$

And the probability of a parse T is the sum of the probabilities of its possible derivations D ;

$$P(T) = \sum_{\{D:D \text{ derives } T\}} P(D) \quad (3)$$

The output of the parser is, in theory, the most probable parse. In practice, there are issues of computational complexity that prevent this from being calculated directly; instead, a Monte-Carlo sample is taken. Furthermore, although the number of subtrees for any given tree increases exponentially with tree-size, it is possible to reduce DOP trees to a stochastically equivalent PCFG which expands linearly with tree-size; for details see Goodman 1996, 2003)

Bod (ibid p.54) reports accuracies of 85% on the ATIS³ corpus for DOP1. However, better results are achieved with later, more sophisticated versions of DOP – the current state of the art is DOP* (Zollman and Sima’an 2005), which selects parses on the basis of the shortest derivation, and only uses probabilities to tie-break if there is more than one shortest derivation; this approach overcomes the problems of statistical consistency and bias which Johnson (1998) pointed out as afflicting DOP1.

2.2 Unsupervised DOP

Unsupervised DOP (UDOP, Bod 2006b, 2007a, 2007b) extends the DOP approach to bootstrap language without recourse to a training corpus of manually annotated language data as a representation of “prior experience”. UDOP expands on DOP’s maximalist, all-subtrees, approach by using all subtrees of all possible (binary) trees to provide the parser with a resource of subtrees for the construction of derivations (which may be stored within a chart in quadratic space: Bod 1998, pp.40-8). Not only is this shown to achieve state-of-the-art results compared to other unsupervised parsing methods (Bod 2006b, 2007a), it is also shown to outperform state-of-the-art *supervised* parsing techniques when evaluated as a language model for a practical application (Machine Translation), rather than using the rather academical and the-

³ Air Transport Information System – part of the Penn Treebank.

ory-laden standard of agreement with the judgements of manual annotations on a corpus (Bod 2007b).

In the UDOP* implementation of UDOP, all subtrees of all possible trees are extracted from the training corpus, and from this exemplar-base, the shortest derivation (in the fashion of DOP*) for each string is calculated (again, the complexity of the task may be reined in using Goodman’s PCFG-reduction, 1996, 2003). The set of trees that results from this is then converted again to subtrees and used as a Stochastic Tree Substitution Grammar, which is then used to parse the test strings. In UML-DOP, another implementation, this last step is iterated over the training data until there is negligible reduction in cross-entropy. Bod (2007a) shows that these methods applied to miniature “toy” corpora can be used to explain linguistically interesting phenomena, such as long-distance agreement and “movement”, as emerging out of simpler structures without themselves being found in either experience or “hardwired” grammar. However, the batch learning methods noted above are cognitively implausible. This is not a criticism of the approach in itself; Batterman (2005) shows that idealisations, however unrealistic in themselves, are necessary in scientific modeling in order to explain universalities which more “realistic” models would miss. However, UDOP cannot itself model the time-course of developmental learning processes, which are by nature incremental. It is to make good this deficit that we have developed the Darwinised DOP (DDOP) approach.

2.3 Darwinised DOP

Darwinised Data-Oriented Parsing is a new unsupervised parsing algorithm which allows the time-course of pattern-learning to be modeled. Unlike previous DOP algorithms, it begins with a completely empty training set; it is fed strings one by one, and its own outputs are entered into the training set. In so doing, it exploits a hitherto underexploited property of exemplar-based systems; when an exemplar (subtree) is reused in producing the system’s eventual output, this output contains a new copy of the reused exemplar which is inserted back into the exemplar-base upon which the algorithm operates; thus, exemplars become *replicators*; packets of information coupled to mechanisms by which new copies of themselves are generated. Furthermore, exemplars which are able to be used more often – those that are more highly *generalisable* – are

likely to make more new copies of themselves; thus we find a *selection pressure* favouring generalisability.⁴ It is also worth noting that because the subsequent sampling of subtrees from a stored tree can and most likely will cross-cut the substitution-sites of the original derivations by which the stored tree was created, replication of subtrees is *recombinant*; exemplars do not just reproduce, they reproduce sexually.

Trees have a limited lifespan, and are erased from memory after K many parses have been generated following their creation. This serves two functions; firstly, if the dataset is small, the training data may be iterated through several times, but because DDOP uses the DOP* shortest derivation method, it must be prevented from seeing strings for which it already has a complete parse in memory, or else it will simply return the parse it gave before, which of course can be generated in a single-step derivation. Secondly, death is an essential component of evolving systems. Without death, maladaptive and primitive forms are allowed to remain in the system, still reproducing, albeit at a slower rate than newer, more highly evolved replicators.

As mentioned above, DDOP uses DOP*’s shortest-derivation method of parsing, but because it begins with an empty exemplar-base, it needs to have a backoff behaviour, in case it encounters a situation, at some point in a derivation, where no subtree can be found in the exemplar-base which would allow the derivation to continue. When a backoff subtree is generated, the following information only is used:

- The node-label l of the substitution site. This does not affect the probability model, but determines the node-label of the root of the new subtree. In fact, this feature is redundant in all the versions of the model tested so far, as the number of available node labels has been limited to one.
- The substring $w_x \dots w_x$, of the total string being parsed, which represents the largest possible substring capable of eventually

⁴ This notion of exemplars as evolving replicators is not without precedent; see Batali 1994; however, Batali’s model concerned the evolution of a shared linguistic code in a population of agents, rather than exploiting this property for the individual learning of a preexisting language. We are also indebted to Kirby’s (1999, for instance) insight that evolving linguistic systems will favour generalisability, though again Kirby’s models concern evolution over glsogenetic, generational time-scales.

being daughters of the node at the substitution site.

We have tested two versions of DDOP, Non-Folding DDOP (NF-DDOP) and Folding DDOP (F-DDOP), each of which use different backoff procedures: RANDOM and FLAT, respectively. The other difference between the two versions of DDOP concerns the interpretation of flat structures in parses (here taken to mean context-free rewrite rules with an arity of three or greater) – internal nodes with three or more immediate daughters. In NF-DDOP, the RANDOM backoff procedure generates subtrees as a random sample of the complete set of all subtrees of all possible parses of substring $w_i \dots w_{i'}$ with root node t . A description of how this is calculated can be found at <http://www.cs.st-andrews.ac.uk/backoff.pdf>. No distinction is made between flat and deep structures. F-DDOP, by contrast, takes flat structures to be an indicator uncertainty regarding actual, lower-arity structure, and therefore as a shorthand for a set of possible lower-arity structures, or “foldings”, wherein the high-arity context free rewrite rules (subtree of depth one) is replaced with a subtree of equal or greater depth, with the same root and frontier. In order to reduce computational load, this set of allowable foldings is limited to subtrees of depth one or two, in the case of the depth two foldings further limited to foldings containing no more than one internal node⁵. By way of example, figure 3 shows all the allowable foldings of a 4-ary subtree.

Derivations proceed one substitution at a time. DOP* subtrees extracted from the training data are always used provided there exists at least one which fits the string being parsed. If at any point no such subtree can be found, then a backoff subtree is generated at that step only. A Monte Carlo sample of N derivations is generated (in the simulations reported here, $N = 1500$). In order to introduce an element of mutation to the process (crucial in any evolving system), a single derivation consisting only of backoff subtrees is occasionally added to the sample, at a probability given by $p(\text{AllBackoff})$, which is a parameter set before the run commences, where $0 \leq p(\text{AllBackoff}) \leq 1$. Following the procedure of

⁵ Note that although this limitation excludes possible parses which may be linguistically pertinent from the immediate folding event (in the case of structures of arity greater than 3), any flat structures which remain can, when re-used, be folded again, allowing the entire space of possible folding to be explored eventually.

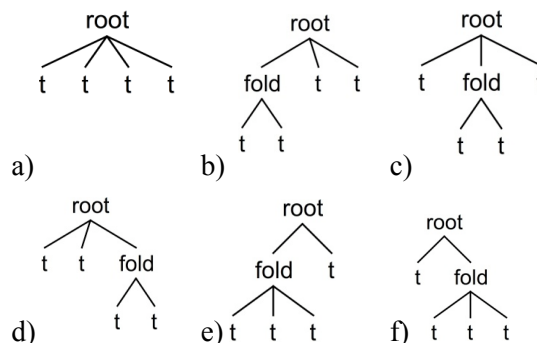


Figure 3: 3(a-f) show the allowable foldings of the 4-ary subtree in 3(a).

DOP*, the shortest derivation in the sample is selected as the output parse. If there is more than one parse with the shortest derivation, DOP1 probabilities are applied to tie-break, following equations 1-3 above. The chosen parse is added to the exemplar-base. Note that because, if an all-backoff derivation is included in the sample, it is assessed for derivation-length and probability just like all the rest, this mutation procedure can only actually introduce novel structures if the random derivation is shorter than all the others in the sample, because only then does it avoid the probability-based tiebreak, which precisely penalizes novelty and favours well-known structures.

Initially, because the exemplar-base is empty, the backoff behaviour is the only behaviour, which, over time, gives way to mostly using corpus trees. Importantly, whenever the outputted tree is derived from subtrees taken from memory, the output tree contains new copies of all those subtrees. In this way, subtrees replicate and more generalisable subtrees are selected.

3 Tests & Results

3.1 Test 1: Six-line toy corpus

The first test each of the versions of DDOP were subjected to was a very simple toy corpus, consisting of six three-word sentences;

Corpus	Oracle
<i>The dog barks.</i>	[<i>The dog</i>] barks.
<i>Watch the dog.</i>	Watch [<i>the dog</i> .]
<i>The dog eats.</i>	[<i>The dog</i>] eats.
<i>The cat barks.</i>	[<i>The cat</i>] barks.
<i>Watch the cat.</i>	Watch [<i>the cat</i> .]
<i>The cat eats.</i>	[<i>The cat</i>] eats.

Table 1: six-line corpus

This initial task was simply to recognise that “the cat” and “the dog” are constituents. The advantage of initially testing the models on this very simple toy corpus is that it is very easy to analyse what the model is doing. The toy corpus contains two sentence types, V-NP and NP-V (V and N for short), each of which may be parse as left-branching, flat or right branching (L, F and R, respectively). After the first three parses on its run, the parser always has parses of $n-1$ of the n sentences in memory; thus each step in an iteration through the data, each stored parse exemplifies one of the 3 possible parse-structures, giving 3^{n-1} possible memory-states for each step, and $n(3^{n-1})$ possible memory states overall. Of the memory-states possible at any step, nine represent a consistent assignment of parse-structures (L/F/R) to sentence-types (V/N), and of these only VR-NL represents a successful outcome. Inconsistent states are never stable; in the absence of mutation they are inaccessible and if accessed as a result of mutation the parser will return to a consistent state within one iteration. Barring multiple mutations it is possible to calculate which consistent state that will be. We call the n states of a consistent assignment (one state for each step), plus all the inconsistent states that lead predictably into it, a “territory” within the state-space. If the states of a consistent assignment always predictably lead into states in the same territory at the next step, the territory will form a cyclical trajectory through state-space, and will be stable barring a disruptive mutation; if they do not, they will lead into the territory of another consistent assignment. A mutation while the parser is on a cyclical trajectory will have one of the following outcomes:

- Parser moves into an inconsistent state of the same territory, returns to cycle.
- Parser moves into an inconsistent state of another territory, changes cycle.

For NF-DDOP, the model was first probed using a state-space model of all 108 of the possible memory-states of the parser based on a four-line version of the above corpus, minus the lines “the dog eats” and “the cat eats”. This analysis found that eight of the nine consistent assignments are stable cycles, with VR-NR being the only exception. VR-NR states always run off to the “true” VR-NL assignment. This means that a mutation in the VR-NL cycle has a greater likelihood of being non-disruptive (returning to VR-NL) than a mutation in any other cycle, and a disruptive

mutation in another cycle is more likely to result in a shift to VR-NL than to any other cycle. However, with seven other cyclical territories to compete with, the parser still spends the majority of its time in states other than VR-NL; VR-NL is more robust than the other cyclical modes, but no modes are wholly robust. This finding was first calculated *a priori* based on the state-space model, then confirmed empirically. Tests with the full 6-line corpus found further destabilization of VR-NL (Chart 1 below).

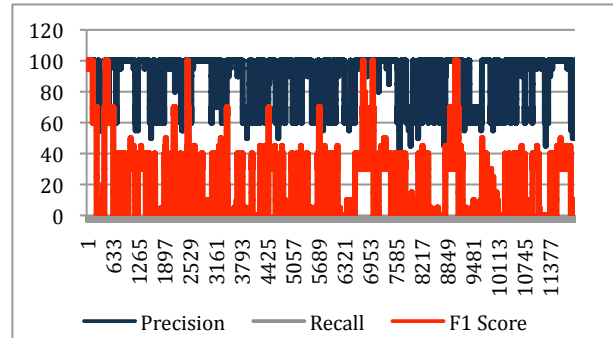


Chart 1: Performance of NF-DDOP on the six-line test corpus

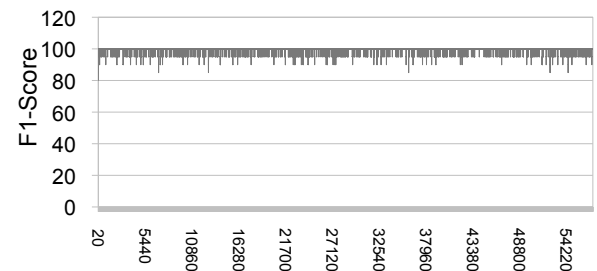


Chart 2: F1-Scores for F-DDOP on the six-line mini-corpus

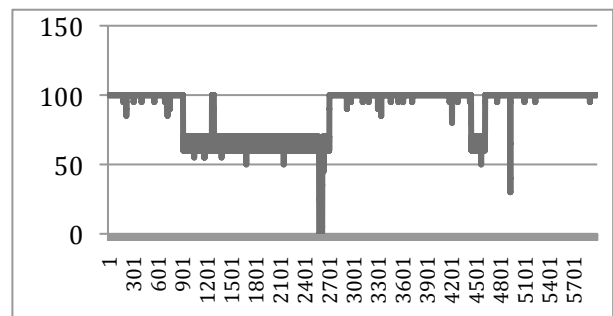


Chart 3: F1 Scores for F-DDOP on the six-line mini-corpus using the RANDOM backoff routine. Each data point averages over the 20 parses, sampled at 10-parse intervals.

In contrast, for F-DDOP, VR-NL is the only accessible territory, and all other consistent assignments run off to it, with the result that VR-NL is completely robust, and the parser’s performance on the six-line toy corpus holds fast to F1-Scores of 100%, as shown in Chart 2 below.

This is because any state containing flat parse-structures runs off to VR-NL, and states contain-

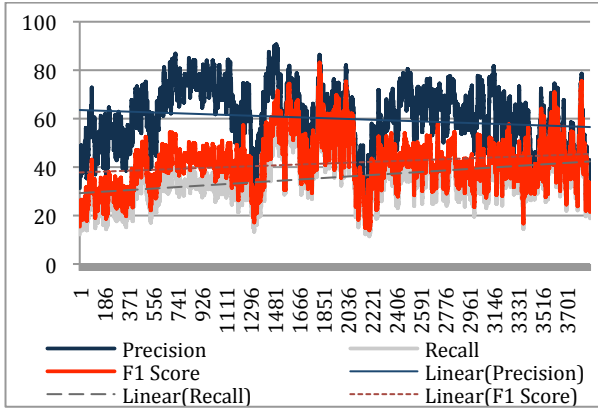


Chart 4: NF-DDOP performance on the 20-line toy corpus.

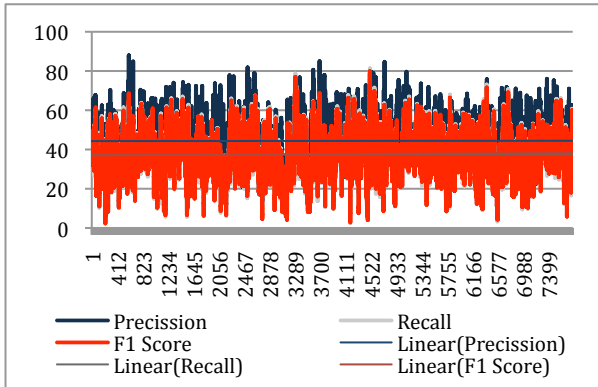


Chart 5: F-DDOP performance on the 20-line toy corpus

ing VL and/or NR are inaccessible, because the FLATback-off always yields VF and NF. This is made clearer by comparison with a hybrid DDOP, in which F-DDOP is combined with the RANDOM back-off routine. Here we find that VL-NL is accessible and cyclical, reducing the robustness of VR-NL.

3.2 Test 2: 20-line toy corpus

The second round of tests used a slightly larger and more complex corpus – this time of 20 sentences, varying between 3 and 11 words in length. Again, notable differences were found between NF-DDOP and F-DDOP. Charts 4 and 5 show the performance of the model on this corpus.

In both cases, the performance of the model was not especially great, with only NF-DDOP showing an overall trend towards improvement over the course of the run; however, the difference in mean F1-score is small; 41.6 for NF-DDOP compared to 37.7 for F-DDOP. Most puzzling of all was the wild instability of F-

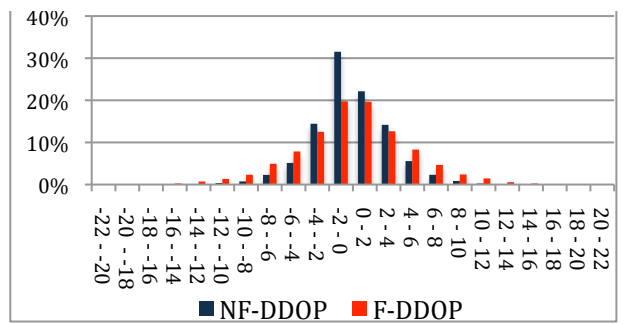


Chart 6: frequency distribution of differences between F1-score at adjacent datapoints in the tests of NF-DDOP and F-DDOP on the 20-line toy corpus.

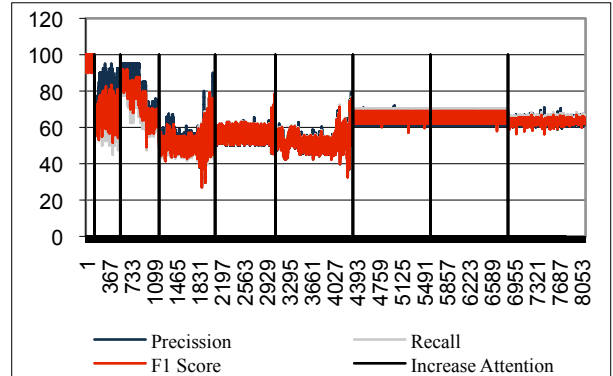


Chart 7: F-DDOP performance with gradually increasing attention span

DDOP on this test, especially when contrasted with its robustness in the simpler test. Chart 6 shows a histogram of the changes in F1-score between adjacent datapoints in the preceding two tests. Note that the values for F-DDOP are rather less sharply peaked and have rather fatter tails. Clearly, further research is required to make sense of this.

More interestingly, a second test was done on F-DDOP in which the model was subject to limitations on attention, which were loosened as time went on: specifically, for the first 700 iterations through the data, it ignored all but the 3-word sentences, then continued in 700-iteration blocks, each with an attention span one word greater than the previous, until it could see the whole corpus. Not only was the overall performance much better, it also showed much greater stability, as can be seen in Chart 7 above

4 Discussion

UDOP has the unrestricted ability to sample the entire range of possible subtrees of possible parses of the entirety of whatever dataset it is presented with, and has a limitless scope to revisit and reconsider past judgements on the basis of new information. As such, we would not expect DDOP to exceed it in power; having access only to a limited subset of the UDOP subtree-set, and strictly limited powers to revisit and reanalyze past judgements, it is fair to assume that at best the performance of UDOP represents a theoretical upper limit of the power of DDOP. As

such, the questions we must attend to are as follows; how far and in what way can we limit the sampling and reanalyzing power of unsupervised DOP systems, and still approach the convergence properties of UDOP in the limit of learning? And if this is possible, can the restrictions adduced be illuminatingly related to empirical work in Developmental and Evolutionary Linguistics?

DDOP in its original form underperformed because it was an unsystematic search through too large a parameter space for too small an error minimum. The population of exemplars is not best conceptualized as an assembly of individuals in a species, who, in competing with their conspecifics, enhance the genetic health of the whole; rather, they are best seen as an assembly of species (though, like bacteria, promiscuously engaged in lateral gene transfer), competing for space in a shifting landscape of ecological niches defined by statistical patterns in the language data. From the viewpoint of an individual exemplar, the ecology of this landscape is as much defined by its competitors and potential mates as by the language-data itself. The problem is the selfishness of replicators described by Dawkins (1976, Williams 1966) in the *Selfish Gene*. Evolution is a dumb, blind process, and selection concerns only the individual replicator in its immediate fitness landscape.

However two modifications to the original DDOP algorithm, which can both be independently motivated in terms of cognitive realism, have been shown to usefully constrain the availability of niches so as to produce more favourable results.

The first, the introduction of the Folding procedure changes the nature of the algorithm's initial assumptions regarding unknown structure; rather than assigning an arbitrary structure to everything on first sight, no initial structure is assumed, allowing the learner to pick recurring motifs out of the data as they are spotted. (see Saffran, Aslin and Newport 1996 on the rapidity with which infants can pick out motifs). The analogue of this behaviour, in terms of evolutionary biology, is phenotypic plasticity, whereby the genotype of a species provides for multiple possible developmental pathways, to multiple phenotypes, modulated in selectionally advantageous ways by environmental conditions. (West-Eberhard 2003). We may consider the case of higher-arity structures in F-DDOP to be the only instance in DDOP of a separation between genotype and phenotype; in the case of the ternary structures in the six-line corpus, the underlying

flat structure is analogous to the genotype, while the flat, left-branching and right-branching subtrees extracted from it in subsequent derivations may be understood as alternative phenotypes. This sort of adaptive plasticity allows "populations to move into new adaptive zones without abandoning old ones ... [a]lternative phenotypes enable condition sensitive evolutionary experimentation within populations" (West-Eberhard 2003, p.392). These variations in phenotype arise out of an interaction between "genotype" and environment (DeWitt and Scheiner 2004b) – understood as both the incoming stream of language inputs and the accumulated store of other exemplars – in a manner that responds adaptively to the frequencies of linguistic patterns in the environment.

Secondly, we introduced the assumption of an initial limitation on the length of sentences the parser is exposed to, which expands with maturation. It is known that adults' speech to infants tends to be characterized by shorter, simpler utterances than normal adult speech (Cameron-Faulkner, Lieven and Tomasello 2003); furthermore, children's speech is characterized by a gradual increase in Mean Length of Utterance (MLU: calculated as morphemes-per-utterance) with maturation (Brown 1973: p270-5), and it has been shown that MLU is a better predictor of comprehension of syntactically complex adult utterances than chronological age (de Villiers and de Villiers, 1973; see also Elman 1993 for further computational work on the payoff of "starting small"). This not only improved overall performance substantially, but also achieved a much-needed gain in stability.

5 Conclusion

We have seen that, by building in additional cognitively realistic assumptions, the overall performance of DDOP, both in terms of average parse quality, and diachronic stability is considerably enhanced; this in itself should be taken as a *prima facie* indication of DDOP's promise as a platform for developmental cognitive modeling. We also note with interest a possible resonance between the current model and neural models of development as evolution at the ontogenetic scale: Edelman's (1987) "Neural Darwinism", or more saliently Fernando, Karishma and Szathmáry's (2008) work on neural-developmental evolution with true replication. Future research in DDOP will investigate the role of the global properties of the exemplar base in determining

the evolutionary dynamics in relation to which exemplars compete and die, and the success or failure of Data-Oriented models of learning and cognition.

References

- Batali, J. 1994. "Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax". In R. Brooks and P. Maes, editors, *Artificial Life IV*. 160-171. Cambridge, MA: MIT Press.
- Batterman, R., 2005. "Critical Phenomena and Breaking Drops: Infinite Idealizations in Physics", *Studies In History and Philosophy of Modern Physics*, 36:225-244.
- Bod, R., 1992. A Computational Model of Language Performance: Data-Oriented Parsing. In *Proc. COLING 1992*, 855--859. Stroudsburg: Association for Computational Linguistics.
- Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*. Stanford: CSLI Publications.
- Bod, R., 2006a. Exemplar-Based Syntax: How to Get Productivity from Examples. *The Linguistic Review* 23, 291-320.
- Bod, R., 2006b. An All-Subtrees Approach to Unsupervised Parsing. In *Proc. ACL 2006*, 865-872. Stroudsburg: Association for Computational Linguistics.
- Bod, R., 2007a. Is the End of Supervised Parsing in Sight? In *Proc. ACL 2007*, 400-407. Stroudsburg: Association for Computational Linguistics.
- Bod, R., 2007b. A Linguistic Investigation into U-DOP. In *Proc. Workshop on Cognitive Aspects of Computational Language Acquisition ACL 2007*, 1-8. Stroudsburg: Association for Computational Linguistics.
- Bod, R., 2008. From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. Accepted for publication in *Cognitive Science*.
- Bod, R., R. Scha and K. Sima'an, (eds.), 2003. *Data-Oriented Parsing*. Stanford, CA: Centre for the Study of Language and Information.
- Borensztajn, G., Zuidema, W., & Bod, R., 2008. Children's grammars grow more abstract with age: evidence from an automatic procedure for identifying the productive units of language. In *Proc. Cog-Sci 2008*, 47-51. Austin: Cognitive Science Society.
- Brown, R., 1973. *A First Language: The Early Stages*. Cambridge, Mass.: Harvard University Press.
- Cameron-Faulkner, T., E. Lieven and M. Tomasello, 2003. A construction-based analysis of child directed speech. *Cognitive Science* 23, 843-873.
- Dawkins, R., 1976. *The Selfish Gene*, Oxford: Oxford University Press.
- De Villiers, J. and P. de Villiers, 1973. Development of the Use of Word Order in Comprehension, *Journal of Psycholinguistic Research*, 2; 331-341.
- DeWitt, T, and S. Scheiner, 2004a. *Phenotypic Plasticity: Functional and Conceptual Approaches*. Oxford: Oxford University Press.
- DeWitt, T. and S. Scheiner, 2004b. "Phenotypic Variation from Single Genotypes: A Primer", in DeWitt and Scheiner 2004a, 1-9.
- Edelman, G. 1987. *The Theory of Neuronal Group Selection*. New York NY: Basic Books.
- Elman, J. 1993. "Learning and development in neural networks: The importance of starting small". *Cognition*, 48(1):71-99.
- Fernando C, K. Karishma and E. Szathmáry. 2008. "Copying and Evolution of Neuronal Topology". *PLoS ONE* 3(11): e3775.
- Goodman, J., 1996. Efficient algorithms for parsing the DOP model. In *Proc. EMNLP 1996*, 143-152. Stroudsburg: Association for Computational Linguistics
- Goodman, J., 2003. "Efficient parsing of DOP with PCFG-reductions". In Bod, Scha and Sima'an 2003, 125-146.
- Johnson, M. 1998. "The DOP Estimation Method is Biased and Inconsistent", *Computational Linguistics*, 28:71-76.
- Kirby, S. 1999. "Learning, bottlenecks, and infinity: a working model of the evolution of syntactic communication", *Proc. AISB'99 Symposium on Imitation in Animals and Artifacts*, 55-63. Society of the Study of Artificial Intelligence and the Simulation of Behaviour.
- Manning, C. and H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass; The MIT Press.
- Saffran, J., R. Aslin, and E. Newport. 1996. "Statistical learning by 8-month-old infants". *Science*, 274:1926-1928.
- Scha, R., 1990. "Taaltheorie en Taaltechnologie: Competence en Performance", in Q. de Kort and G. Leerdam (eds.), *Computertoepassingen in de Neerlandistiek*, Almere: Landelijke Vereniging van Neerlandici.
- West-Eberhard, M, 2003. *Developmental Plasticity and Evolution*, Oxford: Oxford University Press.
- Williams, G., 1966. *Adaptation and Natural Selection*. Princeton: Princeton University Press.
- Zollmann, A., and K. Sima'an, 2005. A consistent and efficient estimator for data-oriented pars-

ing. *Journal of Automata, Languages and Combinatorics* 10, 367-388.

Language Diversity across the Consonant Inventories: A Study in the Framework of Complex Networks

Monojit Choudhury

Microsoft Research India, Bangalore, India – 560080

Email: monojitc@microsoft.com

Animesh Mukherjee, Anupam Basu and Niloy Ganguly

Indian Institute of Technology, Kharagpur, India – 721302

Ashish Garg and Vaibhav Jalan

Malaviya National Institute of Technology, Jaipur, India – 302017

Abstract

In this paper, we attempt to explain the emergence of the linguistic diversity that exists across the consonant inventories of some of the major language families of the world through a complex network based growth model. There is only a single parameter for this model that is meant to introduce a small amount of randomness in the otherwise preferential attachment based growth process. The experiments with this model parameter indicates that the choice of consonants among the languages within a family are far more preferential than it is across the families. Furthermore, our observations indicate that this parameter might bear a correlation with the period of existence of the language families under investigation. These findings lead us to argue that preferential attachment seems to be an appropriate high level abstraction for language acquisition and change.

1 Introduction

In one of their seminal papers (Hauser et al., 2002), Noam Chomsky and his co-authors remarked that if a Martian ever graced our planet then it would be awe-struck by the unique ability of the humans to communicate among themselves through the medium of language. However, if our Martian naturalist were meticulous then it might also note the surprising co-existence of 6700 such mutually unintelligible languages across the world. Till date, the terrestrial scientists have no definitive answer as to why this linguistic diversity exists (Pinker, 1994). Previous work in

the area of language evolution has tried to explain the emergence of this diversity through two different background models. The first one assumes that there is a set of predefined language configurations and the movement of a particular language on this landscape is no more than a random walk (Tomlin, 1986; Dryer, 1992). The second line of research attempts to relate the ecological, cultural and demographic parameters with the linguistic parameters responsible for this diversity (Arita and Taylor, 1996; Kirby, 1998; Livingstone and Fyfe, 1999; Nettle, 1999). From the above studies, it turns out that linguistic diversity is an outcome of the language dynamics in terms of its evolution, acquisition and change.

In this work, we attempt to investigate the diversity that exists across the consonant inventories of the world's languages through an evolutionary framework based on network growth. The use of a network based model is motivated from the fact that in the recent years, complex networks have proved to be an extremely suitable framework for modeling and studying the structure and dynamics of linguistic systems (Cancho and Solé, 2001; Dorogovtsev and Mendes, 2001; Cancho and Solé, 2004; Solé et al., 2005).

Along the lines of the study presented in (Choudhury et al., 2006), we model the structure of the inventories through a *bipartite* network, which has two different sets of nodes, one labeled by the languages and the other by the consonants. Edges run in between these two sets depending on whether a particular consonant is found in a particular language. This network is termed the **Phoneme–Language Network** or **PlaNet** in (Choudhury et al., 2006). We construct five such networks that respectively represent the consonant inventories belonging to the five ma-

major language families namely, the Indo-European (IE-PlaNet), the Afro-Asiatic (AA-PlaNet), the Niger-Congo (NC-PlaNet), the Austronesian (AN-PlaNet) and the Sino-Tibetan (ST-PlaNet).

The emergence of the distribution of occurrence of the consonants across the languages of a family can be explained through a growth model for the PlaNet representing the family. We employ the *preferential attachment* based growth model introduced in (Choudhury et al., 2006) and later analytically solved in (Peruani et al., 2007) to explain this emergence for each of the five families. The model involves a single parameter that is essentially meant to introduce randomness in the otherwise predominantly preferential growth process. We observe that if we combine the inventories for all the families together and then attempt to fit this new data with our model, the value of the parameter is significantly different from that of the individual families. This indicates that the dynamics within the families is quite different from that across them. There are possibly two factors that regulate this dynamics: the innate preference of the speakers towards acquiring certain linguistic structures over others and shared ancestry of the languages within a family.

The prime contribution of this paper lies in the mathematical model that naturally captures and quantifies the diversification process of the language inventories. This diversification, which is arguably an effect of language acquisition and change, can be viewed as a manifestation of the process of preferential attachment at a higher level of abstraction.

The rest of the paper is laid out as follows. Section 2 states the definition of PlaNet, briefly describes the data source and outlines the construction procedure for the five networks. In section 3 we review the growth model for the networks. The experiments and the results are explained in the next section. Section 5 concludes the paper by explaining how preferential attachment could possibly model the phenomena of language acquisition, change and evolution.

2 Definition and Construction of the Networks

In this section, we revisit the definition of PlaNet, discuss briefly about the data source, and explain how we constructed the networks for each of the families.

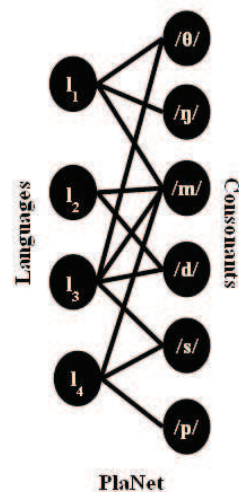


Figure 1: Illustration of the nodes and edges of PlaNet.

2.1 Definition of PlaNet

PlaNet is a bipartite graph $G = \langle V_L, V_C, E_{pl} \rangle$ consisting of two sets of nodes namely, V_L (labeled by the languages) and V_C (labeled by the consonants); E_{pl} is the set of edges running between V_L and V_C . There is an edge $e \in E_{pl}$ from a node $v_l \in V_L$ to a node $v_c \in V_C$ iff the consonant c is present in the inventory of the language l . Figure 1 illustrates the nodes and edges of PlaNet.

2.2 Data Source

We use the UCLA Phonological Segment Inventory Database (UPSID) (Maddieson, 1984) as the source of data for this work. The choice of this database is motivated by a large number of typological studies (Lindblom and Maddieson, 1988; Ladefoged and Maddieson, 1996; de Boer, 2000; Hinskens and Weijer, 2003) that have been carried out on it by earlier researchers. It is a well known fact that UPSID suffers from several problems, especially those involving representational issues (Vaux and Samuels, 2005). Therefore, any analysis carried on UPSID and the inferences drawn from them are subject to questions. However, the current analysis requires a large amount of segment inventory data and to the best of our knowledge UPSID is the biggest database of this kind. Moreover, we would like to emphasize that the prime contribution of this work lies in the mathematical modeling of the data rather than the results obtained, which, as we shall see shortly, are not very surprising or novel. The current model applied to a different database of segment inven-

tories may lead to different results, though we believe that the basic trends will remain similar. In essence, the results described here should be taken as indicative and not sacrosanct.

There are 317 languages in the database with 541 consonants found across them. From these data we manually sort the languages into five groups representing the five families. Note that we included a language in any group if and only if we could find a direct evidence of its presence in the corresponding family. A brief description of each of these groups and languages found within them are listed below (Haspelmath et al., 2005; Gordon, 2005).

Indo-European: This family includes most of the major languages of Europe and south, central and south-west Asia. Currently, it has around 3 billion native speakers, which is largest among all the recognized families of languages in the world. The total number of languages appearing in this family is 449. The earliest evidences of the Indo-European languages have been found to date 4000 years back.

Languages – Albanian, Lithuanian, Breton, Irish, German, Norwegian, Greek, Bengali, Hindi-Urdu, Kashmiri, Sinhalese, Farsi, Kurdish, Pashto, French, Romanian, Spanish, Russian, Bulgarian.

Afro-Asiatic: Afro-Asiatic languages have about 200 million native speakers spread over north, east, west, central and south-west Africa. This family is divided into five subgroups with a total of 375 languages. The proto-language of this family began to diverge into separate branches approximately 6000 years ago.

Languages – Shilha, Margi, Angas, Dera, Hausa, Kanakuru, Ngizim, Awiya, Somali, Iraqw, Dizi, Kefa, Kullo, Hamar, Arabic, Amharic, Socotri.

Niger-Congo: The majority of the languages that belong to this family are found in the sub-Saharan parts of Africa. The number of native speakers is around 300 million and the total number of languages is 1514. This family descends from a proto-language, which dates back 5000 years.

Languages – Diola, Temne, Wolof, Akan, Amo, Bariba, Beembe, Birom, Cham, Dagbani, Doayo, Efik, Ga, Gbeya, Igbo, Ik, Koma, Lelemi, Senadi, Tampulma, Tarok, Teke, Zande, Zulu, Kadugli, Moro, Bisa, Dan, Bambara, Kpelle.

Austronesian: The languages of the Austronesian family are widely dispersed throughout the islands of south-east Asia and the Pacific. There are 1268

Networks	$ V_L $	$ V_C $	$ E_{pl} $
IE-PlaNet	19	148	534
AA-PlaNet	17	123	453
NC-PlaNet	30	135	692
AN-PlaNet	12	82	221
ST-PlaNet	9	71	201

Table 1: Number of nodes and edges in the five bipartite networks corresponding to the five families.

languages in this family, which are spoken by a population of 6 million native speakers. Around 4000 years back it separated out from its ancestral branch.

Languages – Rukai, Tsou, Hawaiian, Iai, Adzera, Kaliai, Roro, Malagasy, Chamorro, Tagalog, Batak, Javanese.

Sino-Tibetan: Most of the languages in this family are distributed over the entire east Asia. With a population of around 2 billion native speakers it ranks second after Indo-European. The total number of languages in this family is 403. Some of the first evidences of this family can be traced 6000 years back.

Languages – Hakka, Mandarin, Taishan, Jingpho, Ao, Karen, Burmese, Lahu, Dafla.

2.3 Construction of the Networks

We use the consonant inventories of the languages enlisted above to construct the five bipartite networks – IE-PlaNet, AA-PlaNet, NC-PlaNet, AN-PlaNet and ST-PlaNet. The number of nodes and edges in each of these networks are noted in Table 1.

3 The Growth Model for the Networks

As mentioned earlier, we employ the growth model introduced in (Choudhury et al., 2006) and later (approximately) solved in (Peruani et al., 2007) to explain the emergence of the *degree distribution* of the consonant nodes for the five bipartite networks. For the purpose of readability, we briefly summarize the idea below.

Degree Distribution: The degree of a node v , denoted by k , is the number of edges incident on v . The degree distribution is the fraction of nodes p_k that have a degree equal to k (Newman, 2003). The cumulative degree distribution P_k is the fraction of nodes having degree greater than or equal to k . Therefore, if there are N nodes in a network

then,

$$P_k = \sum_{k=k'}^N p_{k'} \quad (1)$$

Model Description: The model assumes that the size of the consonant inventories (i.e., the degree of the language nodes in PlaNet) are known *a priori*.

Let the degree of a language node $L_i \in V_L$ be denoted by d_i (i.e., d_i refers to the inventory size of the language L_i in UPSID). The consonant nodes in V_C are assumed to be unlabeled, i.e., they are not marked by the articulatory/acoustic features (see (Trubetzkoy, 1931) for further reference) that characterize them. In other words, the model does not take into account the phonetic similarity among the segments. The nodes L_1 through L_{317} are sorted in the ascending order of their degrees. At each time step a node L_j , chosen in order, preferentially gets connected to d_j distinct nodes (call each such node C) of the set V_C . The probability $Pr(C)$ with which the node L_j gets connected to the node C is given by,

$$Pr(C) = \frac{k + \epsilon}{\sum_{\forall C'} (k' + \epsilon)} \quad (2)$$

where k is the current degree of the node C , C' represents the nodes in V_C that are not already connected to L_j and ϵ is the model parameter that is meant to introduce a small amount of randomness into the growth process. The above steps are repeated until all the language nodes $L_j \in V_L$ get connected to d_j consonant nodes.

Intuitively, the model works as follows: If a consonant is very frequently found in the inventories of the languages, then there is a higher chance of that consonant being included in the inventory of a “new language”. Here the term “new language” can be interpreted either as a new and hitherto unseen sample from the universal set of languages, or the formation of a new language due to some form of language change. The parameter ϵ on the other hand ensures that the consonants which are found in none of the languages from the current sample also have a chance of being included in the new language. It is similar to the add- α smoothing used to avoid zero probabilities while estimating probability distributions. It is easy to see that for very large values of ϵ the frequency factor will play a very minor role and the consonants will be chosen randomly by the new language, irrespective of its present prevalence. It

is natural to ask why and how this particular process would model the growth of the language inventories. We defer this question until the last section of the paper, and instead focus on some empirical studies to see if the model can really explain the observed data.

Peruani et al. (2007) analytically derived an approximate expression for the degree distribution of the consonant nodes for this model. Let the average consonant inventory size be denoted by μ and the number of consonant nodes be N . The solution obtained in (Peruani et al., 2007) is based on the assumption that at each time step t , a language node gets attached to μ consonant nodes, following the distribution $Pr(C)$. Under the above assumptions, the degree distribution $p_{k,t}$ for the consonant nodes, obtained by solving the model, is a β -distribution as follows

$$p_{k,t} \simeq A \left(\frac{k}{t}\right)^{\epsilon-1} \left(1 - \frac{k}{t}\right)^{\frac{N\epsilon}{\mu} - \epsilon - 1} \quad (3)$$

where A is a constant term. Using equations 1 and 3 one can easily compute the value of $P_{k,t}$.

There is a subtle point that needs a mention here. The concept of a *time step* is very crucial for a growing network. It might refer to the addition of an edge or a node to the network. While these two concepts coincide when every new node has exactly one edge, there are obvious differences when the new node has degree greater than one. The analysis presented in Peruani et al. (2007) holds good for the case when only one edge is added per time step. However, if the degree of the new node being introduced to the system is much less than N , then Eq. 3 is a good approximation of the emergent degree distribution for the case when a node with more than one edge is added per time step. Therefore, the experiments presented in the next section attempt to fit the degree distribution of the real networks with Eq. 3 by tuning the parameter ϵ .

4 Experiments and Results

In this section, we attempt to fit the degree distribution of the five empirical networks with the expression for $P_{k,t}$ described in the previous section. For all the experiments we set $N = 541$, $t =$ number of languages in the family under investigation and $\mu =$ average degree of the language nodes of the PlaNet representing the family under investigation, that is, the average inventory size for

Network	ϵ for least LSE	Value of LSE
IE-PlaNet	0.055	0.16
AA-PlaNet	0.040	0.24
NC-PlaNet	0.035	0.19
AN-PlaNet	0.030	0.17
ST-PlaNet	0.035	0.03
Combined-PlaNet	0.070	1.47

Table 2: The values of ϵ and the least LSE for the different networks. Combined-PlaNet refers to the network constructed after mixing all the languages from all the families. For all the experiments

the family. Therefore, given the value of k we can compute $p_{k,t}$ using Eq. 3 if ϵ is known, and from $p_{k,t}$ we can further compute $P_{k,t}$. In order to find the best fitting theoretical degree distribution, we vary the value of ϵ in steps of 0.005 within the range of 0 to 1 and choose that ϵ for which the logarithmic standard error¹ (LSE) between the theoretical degree distribution and the empirically observed degree distribution of the real network and the equation is least. LSE is defined as the sum of the square of the difference between the logarithm of the ordinate pairs (say y and y') for which the abscissas are equal. The best fits obtained for each of the five networks are shown in Figure 2. The values of ϵ and the corresponding least LSE for each of them are noted in Table 2. We make the following significant and interesting observations.

Observation I: The very low value of the parameter ϵ indicates that the choice of consonants within the languages of a family is strongly preferential. In this context, ϵ may be thought of as modeling the (accidental) errors or drifts that can occur during language transmission. The fact that the values of ϵ across the four major language families, namely Afro-Asiatic, Niger-Congo, Sino-Tibetan and Austronesian, are comparable indicates that the rate of error propagation is a universal factor that is largely constant across the families. The value of ϵ for IE-PlaNet is slightly higher than the other four families, which might be an effect of higher diversification within the family due to geographical or socio-political factors. Nevertheless, it is still smaller than the ϵ of the Combined-

¹ $LSE = (\log y - \log y')^2$. We use LSE as the goodness of the fit because the degree distributions of PlaNets are highly skewed. There are very few high degree nodes and a large number of low degree nodes. The logarithmic error ensures that even very small errors made while fitting the high degrees are penalized equally as compared to that of the low degrees. Standard error would not capture this fact and declare a fit as good if it is able to replicate the distribution for low degrees, but fits the high degrees poorly.

PlaNet.

The optimal ϵ obtained for Combined-PlaNet is higher than that of all the families (see Table 2), though it is comparable to the Indo-European PlaNet. This points to the fact that the choice of consonants within the languages of a family is far more preferential than it is across the families; this fact is possibly an outcome of shared ancestry. In other words, the inventories of genetically related languages are similar (i.e., they share a lot of consonants) because they have evolved from the same parent language through a series of linguistic changes, and the chances that they use a large number of consonants used by the parent language is naturally high.

Observation II: We observe a very interesting relationship between the approximate age of the language family and the values of ϵ obtained in each case (see Table 3). The only anomaly is the Indo-European branch, which possibly indicates that this might be much older than it is believed to be. In fact, a recent study (Balter, 2003) has shown that the age of this family dates back to 8000 years. If this last argument is assumed to be true then the values of ϵ have a one-to-one correspondence with the approximate period of existence of the language families. As a matter of fact, this correlation can be intuitively justified – the higher is the period of existence of a family, the higher are the chances of transmission errors leading to its diversification into smaller subgroups, and hence, the values of ϵ comes out to be more for the older families. It should be noted that the difference between the values of ϵ for the language families are not significant². Therefore, the aforementioned observation should be interpreted only as an interesting possibility; more experimentation is required for making any stronger claim.

4.1 Control Experiment

How could one be sure that the aforementioned observations are not an obvious outcome of the construction of the PlaNet or some spurious correlations? To this end, we conduct a control experiment where a set of inventories is randomly selected from UPSID to represent a family. The

²Note that in order to obtain the best fit for the cumulative distribution, ϵ has been varied in steps of 0.005. Therefore, the values of ϵ in Table 2 cannot be more accurate than $\epsilon \pm 0.005$. However, in many cases the difference between the best-fit ϵ for two language families is exactly 0.005, which indicates that the difference is not significant.

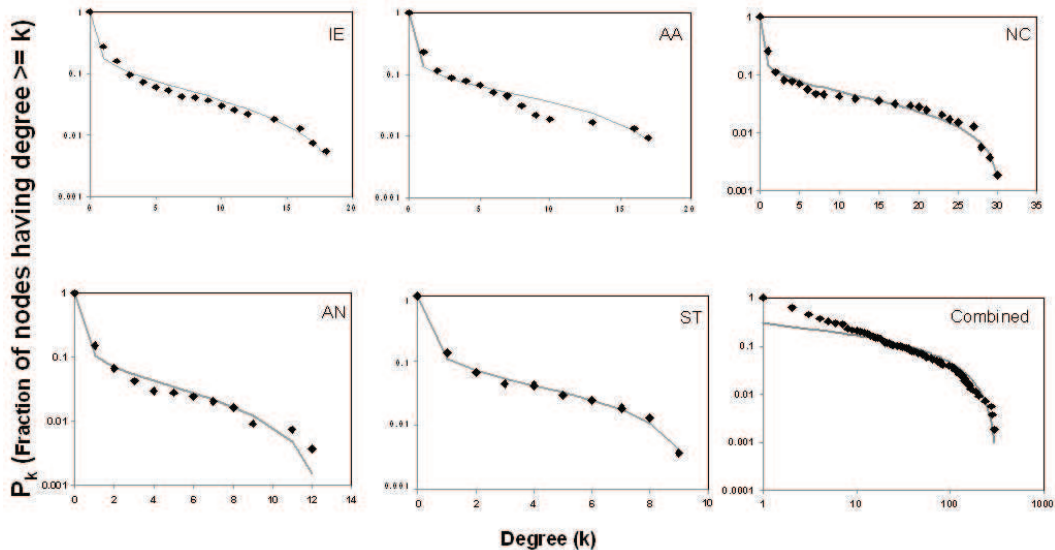


Figure 2: The degree distribution of the different real networks (black dots) along with the fits obtained from the equation for the optimal values of ϵ (grey lines).

Families	Age (in years)	ϵ
Austronasean	4000	0.030
Niger-Congo	5000	0.035
Sino-Tibetan	6000	0.035
Afro-Asiatic	6000	0.040
Indo-European	4000 (or 8000)	0.055

Table 3: Table showing the relationship between the age of a family and the value of ϵ .

number of languages chosen is the same as that of the PlaNets of the various language families. We observe that the average value of ϵ for these randomly constructed PlaNets is 0.068, which, as one would expect, is close to that of the Combined-PlaNet. This reinforces the fact that the inherent proximity among the languages of a real family is not due to chance.

4.2 Correlation between Families

It can be shown theoretically that if we merge two PlaNets (say PlaNet₁ and PlaNet₂) synthesized using the growth model described here using parameters ϵ_1 and ϵ_2 , then the ϵ of the combined PlaNet can be much greater than both ϵ_1 and ϵ_2 when there is a low correlation between the degrees of the consonant nodes between the two PlaNets. This can be understood as follows. Suppose that the consonant /k/ is very frequent (i.e., has a high degree) in PlaNet₁, but the consonant /m/ is not. On the other hand suppose that /m/ is very frequent in PlaNet₂, but /k/ is not. In the combined

PlaNet the degrees of /m/ and /k/ will even out and the degree distribution will therefore, be much less skewed than the original degree distributions of PlaNet₁ and PlaNet₂. This is equivalent to the fact that while ϵ_1 and ϵ_2 were very small, the ϵ of the combined PlaNet is quite high. By the same logic it follows that if the degrees of the consonants are highly correlated in PlaNet₁ and PlaNet₂, then the combined PlaNet will have an ϵ that is comparable in magnitude to ϵ_1 and ϵ_2 . The fact that the ϵ for the Combined-PlaNet is higher than that of family-specific PlaNets, therefore, implies that the correlation between the frequencies of the consonants across language families is not very high.

In order to verify the above observation we estimate the correlation between the frequency of occurrence of the consonants for the different language family pairs (i.e., how the frequencies of the consonants /p/, /t/, /k/, /m/, /n/ ... are correlated across the different families). Table 4 notes the value of this correlation among the five families. The values in Table 4 indicate that, in general, the families are somewhat weakly correlated with each other, the average correlation being ~ 0.47 .

Note that, the correlation between the Afro-Asiatic and the Niger-Congo families is high not only because they share the same African origin, but also due to higher chances of language contacts among their groups of speakers. On the other hand, the Indo-European and the Sino-Tibetan families show least correlation because it is usu-

Families	IE	AA	NC	AN	ST
IE	–	0.49	0.48	0.42	0.25
AA	0.49	–	0.66	0.53	0.43
NC	0.48	0.66	–	0.55	0.37
AN	0.42	0.53	0.55	–	0.50
ST	0.25	0.43	0.37	0.50	–

Table 4: The Pearson’s correlation between the frequency distributions obtained for the family pairs. IE: Indo-European, AA: Afro-Asiatic, NC: Niger-Congo, AN: Austronesian, ST: Sino-Tibetan.

ally believed that they share absolutely no genetic connections. Interestingly, similar trends are observed for the values of the parameter ϵ . If we combine the languages of the Afro-Asiatic and the Niger-Congo families and try to fit the new data then ϵ turns out to be 0.035 while if we do the same for the Indo-European and the Sino-Tibetan families then ϵ is 0.058. For many of the other combinations the value of ϵ and the correlation coefficient have a one-to-one correspondence. However, there are clear exceptions also. For instance, if we combine the Afro-Asiatic and the Indo-European families then the value of ϵ is very low (close to 0.04) although the correlation between them is not very high. The reasons for these exceptions should be interesting and we plan to further explore this issue in future.

5 Conclusion

In this paper, we presented a method of network evolution to capture the emergence of linguistic diversity that manifests in the five major language families of the world. How does the growth model, if at all, captures the process of language dynamics? We argue that preferential attachment is a high level abstraction of language acquisition as well as language change. We sketch out two possible explanations for this fact, both of which are merely speculations at this point and call for detailed experimentation.

It is a well known fact that the process of language acquisition by an individual largely governs the course of language change in a linguistic community. In the initial years of language development every child passes through a stage called *babbling* during which he/she learns to produce non-meaningful sequences of consonants and vowels, some of which are not even used in the language to which they are exposed (Jakobson, 1968; Locke, 1983). Clear preferences can be

observed for learning certain sounds such as plosives and nasals, whereas fricatives and liquids are avoided. In fact, this hierarchy of preference during the babbling stage follows the cross-linguistic frequency distribution of the consonants. This innate frequency dependent preference towards certain phonemes might be because of phonetic reasons (i.e., for articulatory/perceptual benefits). It can be argued that in the current model, this innate preference gets captured through the process of preferential attachment.

An alternative explanation could be conceived of based on the phenomenon of language transmission. Let there be a community of N speakers communicating among themselves by means of only two consonants say $/k/$ and $/g/$. Let the number of $/k/$ speakers be m and that of $/g/$ speakers be n . If we assume that each speaker has l descendants and that language inventories are transmitted with high fidelity then after i generations, the number of $/k/$ speakers should be ml^i and that of $/g/$ speakers should be nl^i . Now if $m > n$ and $l > 1$ then for sufficiently large values of i we have $ml^i \gg nl^i$. Stated differently, the $/k/$ speakers by far outnumbers the $/g/$ speakers after a few generations even though the initial difference between them is quite small. This phenomenon is similar to that of preferential attachment where language communities get attached to, i.e., select consonants that are already highly preferred. In this context ϵ can be thought to model the accidental errors during transmission. Since these errors accumulate over time, this can intuitively explain why older language families have a higher value of ϵ than the younger ones.

In fact, preferential attachment (PA) is a universally observed evolutionary mechanism that is known to shape several physical, biological and socio-economic systems (Newman, 2003). This phenomenon has also been called for to explain various linguistic phenomena (Choudhury and Mukherjee, to appear). We believe that PA also provides a suitable abstraction for the mechanism of language acquisition. Acquisition of vocabulary and growth of the mental lexicon are few examples of PA in language acquisition. This work illustrates another variant of PA applied to explain the structure of consonant inventories and their diversification across the language families.

References

- T. Arita and C. E. Taylor. 1996. A simple model for the evolution of communication. In L. J. Fogel, P. J. Angeline and T. Bäck, editors, *The Fifth Annual Conference On Evolutionary Programming*, 405–410. MIT Press.
- M. Balter. 2003. Early date for the birth of Indo-European languages. *Science* **302**(5650), 1490.
- A.-L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science* **286**, 509–512.
- D. Bickerton. 1990. *Language and Species*, The University of Chicago Press, Chicago.
- B. de Boer. 2000. Self-organization in vowel systems. *Journal of Phonetics*, **28**(4), 441–465.
- R. Ferrer i Cancho and R. V. Solé. 2001. The small-world of human language. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, **268**(1482), 1228–1235.
- R. Ferrer i Cancho and R. V. Solé. 2004. Patterns in syntactic dependency networks. *Phys. Rev. E*, **69**(051915).
- R. G. Gordon (ed.) 2005. *Ethnologue: Languages of the World*, Fifteenth edition, SIL International.
- M. Haspelmath, M. S. Dryer, D. Gil and B. Comrie (ed.) 2005. *World Atlas of Language Structures*, Oxford University Press.
- M. Choudhury, A. Mukherjee, A. Basu and N. Ganguly. 2006. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Main Conference Poster Sessions*, 128–135.
- M. Choudhury and A. Mukherjee. to appear. The structure and dynamics of linguistic networks. In N. Ganguly, A. Deutsch and A. Mukherjee, editors, *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences*, Birkhauser, Springer, Boston.
- S. N. Dorogovtsev and J. F. F. Mendes. 2001. Language as an evolving word web. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, **268**(1485), 2603–2606.
- M. S. Dryer. 1992. The Greenbergian word order correlations. *Language*, **68**, 81–138.
- M. D. Hauser, N. Chomsky and W. T. Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, **298**, 1569–1579.
- F. Hinskens and J. Weijer. 2003. Patterns of segmental modification in consonant inventories: a cross-linguistic study. *Linguistics*, **41**(6), 1041–1084.
- R. Jakobson. 1968. *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.
- H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. L. Barabási. 2000. The large-scale organization of metabolic networks. *Nature*, **406**, 651–654.
- S. Kirby. 1998. Fitness and the selective adaptation of language. In J. R. Hurford, M. Studdert-Kennedy and C. Knight, editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, 359–383. Cambridge: Cambridge University Press.
- P. Ladefoged and I. Maddieson. 1996. *Sounds of the Worlds Languages*, Oxford: Blackwell.
- B. Lindblom and I. Maddieson. 1988. Phonetic universals in consonant systems. In L.M. Hyman and C.N. Li, eds., *Language, Speech, and Mind*, Routledge, London, 62–78.
- D. Livingstone and C. Fyfe. 1999. Modelling the evolution of linguistic diversity. In D. Floreano, J. Nicoud and F. Mondada, editors, *ECAL 99*, 704–708, Berlin: Springer-Verlag.
- J. L. Locke. 1983. *Phonological Acquisition and Change*. Academic Press New York.
- I. Maddieson. 1984. *Patterns of Sounds*, Cambridge University Press, Cambridge.
- D. Nettle. 1999. Is the rate of linguistic change constant? *Lingua*, **108**(2):119–136.
- M. E. J. Newman. 2001. Scientific collaboration networks. *Physical Review E* **64**, 016131.
- M. E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review* **45**, 167–256.
- F. Peruani, M. Choudhury, A. Mukherjee and N. Ganguly. 2007. Emergence of a non-scaling degree distribution in bipartite networks: a numerical and analytical study. *Euro. Phys. Letters* **76**, 28001 (p1–p6).
- S. Pinker. 1994. *The Language Instinct*, New York: William Morrow.
- E. Pulleyblank. 1993. The typology of Indo-European. *Journal of Indo-European Studies*, p. 109.
- José J. Ramasco, S. N. Dorogovtsev, and Romualdo Pastor-Satorras. 2004. Self-organization of collaboration networks. *Physical Review E*, **70**, 036106.
- R. V. Solé, B. C. Murtra, S. Valverde and L. Steels. 2005. Language networks: Their structure, function and evolution. *Santa Fe working paper*, 05-12-042.
- R. Tomlin. 1986. *Basic Word Order: Functional Principles*, Croom Helm, London.
- N. Trubetzkoy. 1931. Die phonologischen systeme. *TCLP* **4**, 96–116.
- B. Vaux and B. Samuel. 2005. Laryngeal markedness and aspiration *Phonology* **22**(3), 96–116.

Author Index

Afantenos, Stergos, 18

Basu, Anupam, 51

Choudhury, Monojit, 51

Clark, Alexander, 26

Cochran, Dave, 42

Cooper, Robin, 1

Dickinson, Markus, 34

Ganguly, Niloy, 51

Garg, Ashish, 51

Hernandez, Nicolas, 18

Jack, Kris, 10

Jalan, Vaibhav, 51

Jochim, Charles, 34

Lappin, Shalom, 26

Larsson, Staffan, 1

Mukherjee, Animesh, 51