A Rule-Based Unsupervised Morphology Learning Framework

Constantine Lignos, Erwin Chan*,

Mitch Marcus, Charles Yang

University of Pennsylvania, *University of Arizona

Morpho Challenge 2009

CLEF 2009, 9/30/2009

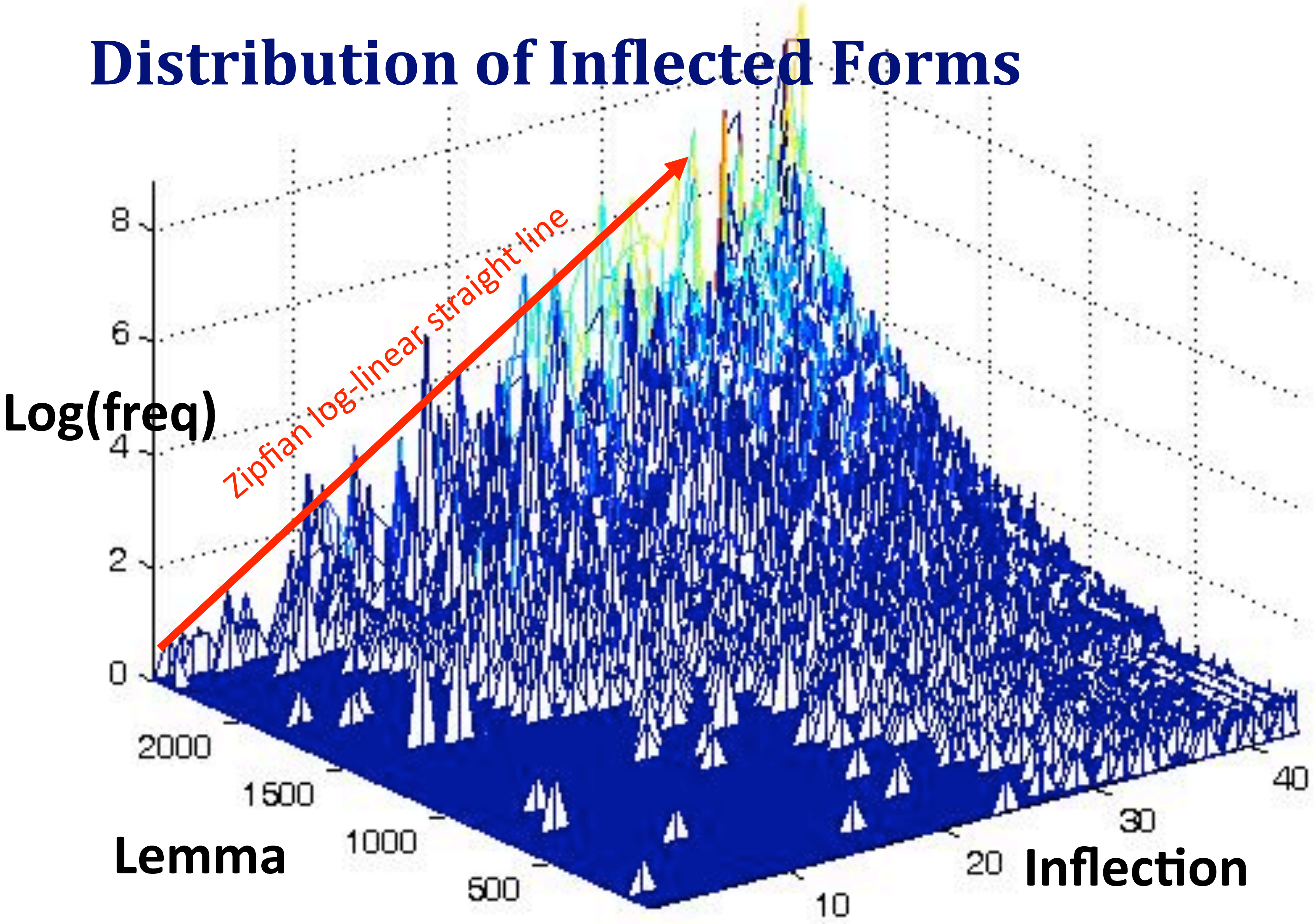
Defining the Task

- **Application of a language acquisition model as a morphological analyzer**
- **How do we define an acquisition model?**
  - Cognitively motivated- the representations it learns are linguistically motivated and cognitively useful
  - Designed for a child's input- Small amounts of sparse data received in an unsupervised fashion
- **Not looking to create a fully psychologically plausible algorithm**
  - While the structures learned are plausible, some parts of the algorithm are computationally expensive for the sake of simplicity

The Learning Model: Chan (2008)

- **Structures and Distributions in Morphology Learning**
- **Provides:**
  - Representation of morphology- Base and Transforms Model
  - Simple bootstrapping algorithm for learning bases and transforms in an unsupervised fashion
- **Enhancements needed for Morpho Challenge:**
  - Adaptation to larger/noisier corpora
  - Morphological analysis output
  - Support for multi-step derivations

Distribution of Inflected Forms



Spanish newswire verbs (2.5 M)

Base and Transforms Model

- Within each syntactic category, the most common inflected form is consistent
- Instead of relying on an abstract stem, we have a "base" form that we can easily identify- the most common inflection in each category
- To model a derived form, apply a transform to a base:

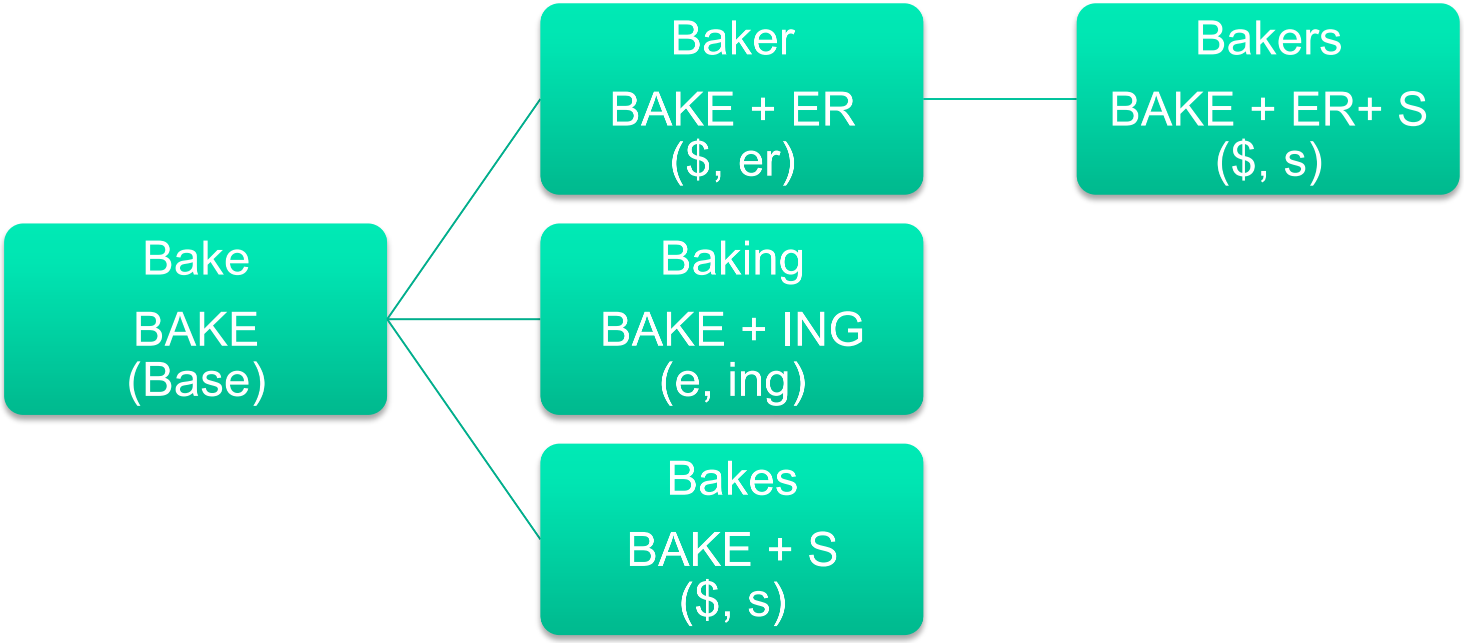
RUN + ($, s) = runs

MAKE + (e, ing) = making

Note: $ is used to represent a null affix

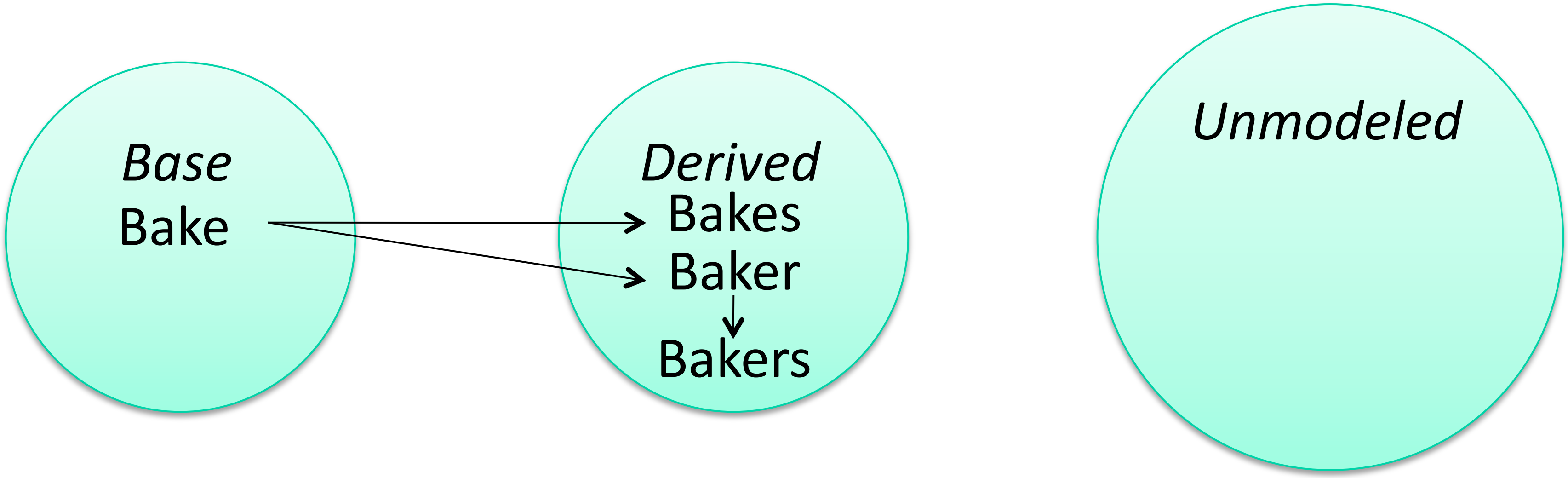
Base and Transforms Model

- **The learner will learn a set of rules (transforms) and the word pairs they apply to (base-derived pairs)**



The Algorithm: Sets

- **A word belongs to one of three sets at any time:**
  - Unmodeled- All words begin in this set
  - Base- Words that are used as a base in a transform and are not derived from anything else
  - Derived- Words that are derived from a base word or another derived word



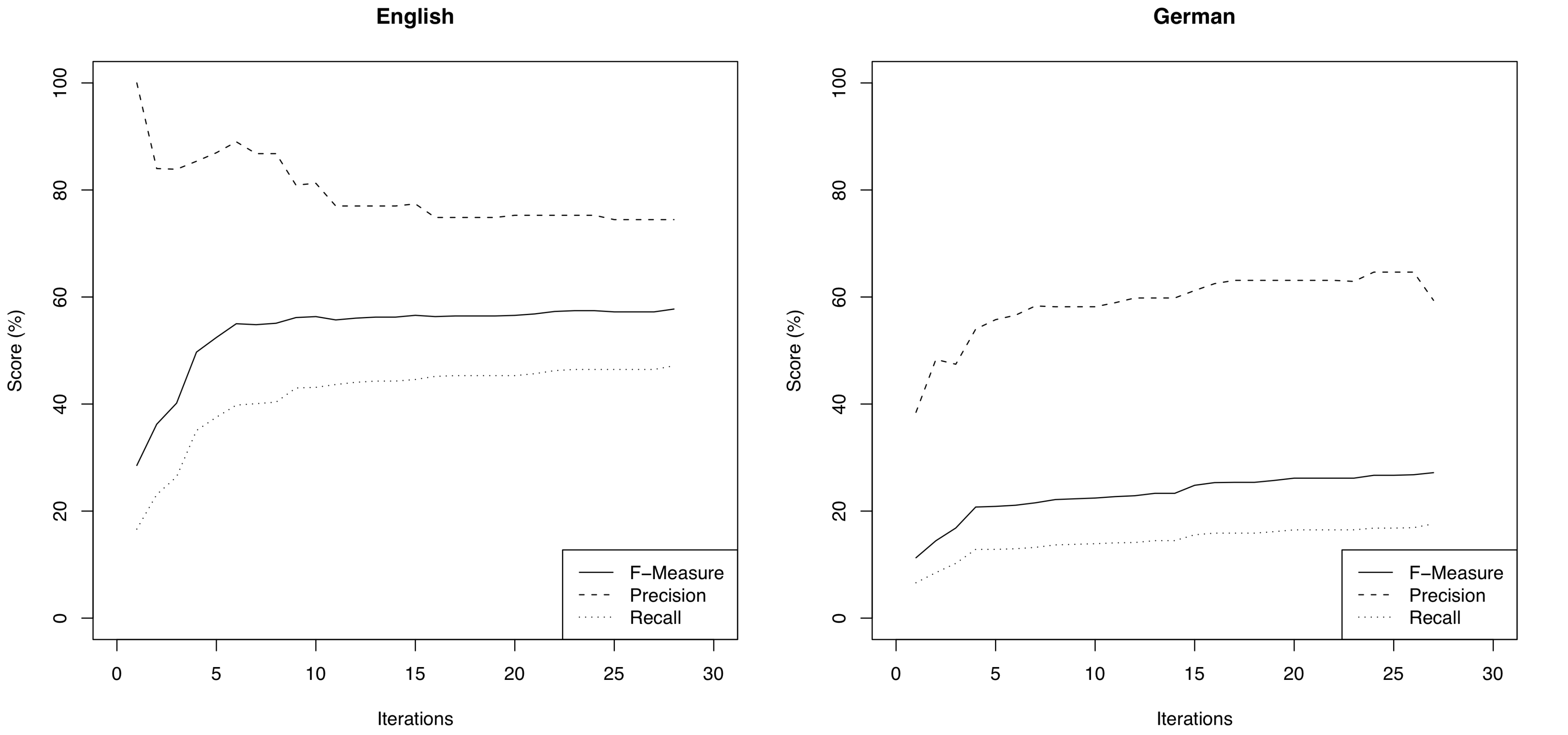
Core Algorithm

1. **Pre-process words and populate the Unmodeled set.**
2. **Until a stopping condition is met, perform the main learning loop:**
    1. Count affixes in words of the (Base + Unmodeled) set and the Unmodeled set.
    2. Hypothesize transforms from words in (Base + Unmodeled) to words in Unmodeled.
    3. Select the best transform.
    4. Reevaluate the words that the selected transform applies to, using the Base, Derived and Unmodeled sets
    5. Move the words used in the transform accordingly.
3. **Break compound words in the Base and Unmodeled sets.**
4. **Output analysis**

English Transforms Learned

	Trans.	Sample Pair
1	+($, s)	scream/screams
2	+($, ed)	splash/splashed
3	+($, ing)	bond/bonding
4	+($, ‘s)	office/office’s
5	+($, ly)	unlawful/unlawfully
6	+(e, ing)	supervise/supervising
7	+(y, ies)	fishery/fisheries
8	+($, es)	skirmish/skirmishes
9	+($, er)	truck/trucker
10	($, un)+	popular/unpopular
11	+($, y)	risk/risky
12	($, dis)+	credit/discredit
13	($, in)+	appropriate/ inappropriate
14	+($, ation)	transform/transformation
15	+($,al)	intention/intentional
16	+(e, tion)	deteriorate/deterioration
17	+(e, ation)	normalize/normalization
18	+(e, y)	subtle/subtly
19	+($, st)	safe/safest
20	($, pre)+	school/preschool
21	+($, ment)	establish/establishment
22	($, inter)+	group/intergroup
23	+(t, ce)	evident/evidence
24	($,se)+	cede/secede
25	+($, a)	helen/helena
26	+(n, st)	lighten/lightest
27	($, be)+	came/became

Performance



Error Types and Proposed Solutions

- **Almost all transforms learned are real morphological rules, although they sometimes have spurious pairs**
  - In English, +($, a) and ($,se)+ are the only spurious transforms out of 27 learned
  - Example spurious pairs for good transforms:
    - gust/disgust
    - pen/penal
    - tent/intent
    - gin/begin
  - Part of the cause is there is no concept of syntactic categories
    - Thus no concept of inflectional/derivational rules
    - Basic approach to category induction in Chan 2008, but needs refinement to identify category of derived forms

Error Types and Proposed Solutions

- **Difficulty learning multistep derivations**
  - Does not predict existence of unseen forms
    - Ex: acidified = ACID + ($, ify) + (y, ied)
    - If *acidify* is not seen in the corpus we won't learn the connection between *acid* and *acidified*
  - The learner needs to understand the productivity of rules in order to decide whether it's likely an unseen form exists
- **Rule representation too simple for other languages**
  - All rules consist of affix changes only
  - Should support wider morphological functions, such as templatic morphology and vowel harmony

Conclusions

- An acquisition model can provide an effective learning framework for a morphological analyzer
- Chan (2008) model and algorithm deliver competitive results in English and German with some adaptation
- To cover more languages, the representations used by the learner needs to be expanded