A database for the accentual patterns in the world's languages

Jeffrey Heinz
University of Delaware
heinz@udel.edu

ICPP 3
National Institute for Japanese Language and Linguistics
December 21, 2013

*This research is supported by NSF award #1123692.

Wilhelm Von Humboldt

"language makes infinite use of finite means"

Wilhelm Von Humboldt

Typology:

1. "Encyclopedia of Types"
2. "Encyclopedia of Categories"

This talk

Encyclopedia of Types

StressTyp2

Encyclopedia of Categories

Computer Science (specifically: a model theoretic approach to formal language theory)

Outline

What is stress and/or accent?

Pintupi (Hansens and Hansen 1969)

a.	*σ́σ*	páɳa	‘earth’
b.	*σ́σσ*	tʲúʈaya	‘many’
c.	*σ́σσ̀σ*	málawàna	‘through from behind’
d.	*σ́σσ̀σσ*	púlinkàlatʲu	‘we (sat) on the hill’
e.	*σ́σσ̀σσ̀σ*	tʲámulìmpatʲùŋku	‘our relation’
f.	*σ́σσ̀σσ̀σσ*	ʈílirìŋulàmpatʲu	‘the fire for our benefit flared up’
g.	*σ́σσ̀σσ̀σσ̀σ*	kúranʲùlulìmpatʲùɻa	‘the first one who is our relation’
h.	*σ́σσ̀σσ̀σσ̀σσ*	yúmaɻìŋkamàratʲùɻaka	‘because of mother-in-law’

What is stress and/or accent?

Latin (Jacobs 1989, Mester 1992, Hayes 1995)

a.	L Ĥ H	a.míː.kus	‘friend, kind’
b.	L H Ĥ H	gu.ber.náː.bunt	‘they will reign’
c.	L L H Ĺ L L	i.ni.miː.kìː.ti.a	‘hostility’
d.	L Ĥ L H	do.més.ti.kus	‘belonging to the house’
e.	Ĥ H	mán.daː	‘entrust (2sg.imp)’
f.	Ĺ H	ká.nis	‘dog’
g.	Ĺ L	hé.ri	‘yesterday’

What is stress and/or accent?

Selkup (Halle and Clements 1983, Idsardi 1992, Walker 2000)

a.	L L L Ĺ	[pynakɨsə́ː]	‘giant!’
b.	L L Ĺ L	[ilɨsóːmɨt]	‘we lived’
c.	Ĺ L L	[qóːkɨtɨlʲ]	‘deaf’
d.	L H L Ĺ	[qumoːqlılíː]	‘your two friends’
e.	H Ĺ L	[uːcóːmɨt]	‘we work’
f.	H L Ĺ L	[uːcɨkkóːqı]	‘they two are working’
g.	Ĺ L	[qúmmɨn]	‘human being’ (gen.)
h.	Ĺ L L	[ámɨrna]	‘eats’
i.	Ĺ L L L	[qólʲcɨmpatɨ]	‘found’

Examples of Generalizations

Pintupi

Primary stress falls on the first syllable and secondary stress on all nonfinal odd syllables.

Latin

Primary stress falls on penultimate syllable if it is heavy else it falls on the antepenult (if there is one) else the penult.

Selkup

Primary stress falls on rightmost heavy syllable. If there are no heavy syllables it fall on the leftmost syllable.

Questions about stress

1. Is stress predictable? In what way?
2. What are the phonetic correlates of stress?
3. How is stress affected by morpho-syntax?
4. How does stress interact with the phonology?

Outline

Collaborators

Rob Goedemans (Leiden University)
Harry van der Hulst (University of Connecticut)

Graduate Research Assistants

@ Delaware	**@ UConn**
Gordon Hemsley	Mary Goodrich
Adam Jardine	Aida Talic
Amanda Payne	

Universiteit Leiden

What is StressTyp2?

- StressTyp2 (ST2) is an international colloborative project to collect and organize the stress, accentual and rhythmic patterns of the world's languages supported by the United States National Science Foundation.

Goals

StressTyp2s purpose is to provide a tool for both researchers and the general public to better understand the nature of stress and accent in the worlds languages.

Problems and Questions

1. Given the variety of linguistic descriptions, how can they be uniformly encoded into a database?
2. Since sources vary in the degree of detail, how can the quality of description be encoded?
3. How can exceptions, and patterned exceptions, be included?
4. How will different linguistic descriptions of the same language be addressed?

Some history

ST2 contains information

- from the original StressTyp (Goedemans et al. 1996; Goedemans and van der Hulst 2009, 2010, inter alia)
- from the Stress Pattern Database (Heinz 2007), which itself was based on The Stress System Database (SSD, Bailey 1995), Hyman's 1977 collection, and Gordon's 2002 typology.
- on over 700 languages, with nearly every language family represented.

Types of Information

1. The focus has been on predictable dominant stress patterns.
2. Some information on subordinate and exceptional stress patterns.
3. Some information about syllable structure as it relates to stress.
4. Some information about morpho-syntax (e.g. compound stress).

Features of StressTyp2

Key Features

1. Transparency
2. Robustness
3. Accessibility
4. Replicability
5. Flexibility
6. Extensibility

1. Transparency

- The source of each piece of information in the database is documented.
- ST2 aims not to impose the views of its designers, but rather to provide a key to the scientific, linguistic literature.

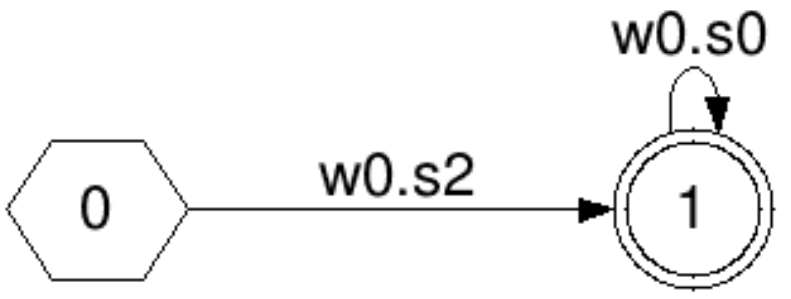
2. Robustness

- The metrical and accentual patterns themselves are described in multiple formats.
- These formats include
  - Linguistic parameter settings
  - The original StressTyp codes
  - The Stress System Database's Syllable Priority Codes
  - Finite-state representations
  - and will soon include English prose

Example: Koromfe

Initial Stress

- STC code: I
- SPC code: 1L
- Linguistic Parameters: Left, Trochaic
- Finite-state diagram:



- English Prose: *Primary stress falls on the initial syllable. There is no secondary stress.*

3. Accessibility

ST2 is freely accessible online for scholars and the public.

st2.ullet.net

This website, while not yet officially announced, is live.

Browsing by language, lect or pattern

- Lects list patterns, attributes, syllabic information, and example words.
- Patterns list their status (dominant, subordinate, exceptional), attributes, theoretical analyses, computational analyses, prose analyses, and other lects with the same pattern.
- Familial and geographical information is also included.

Searching the web interface

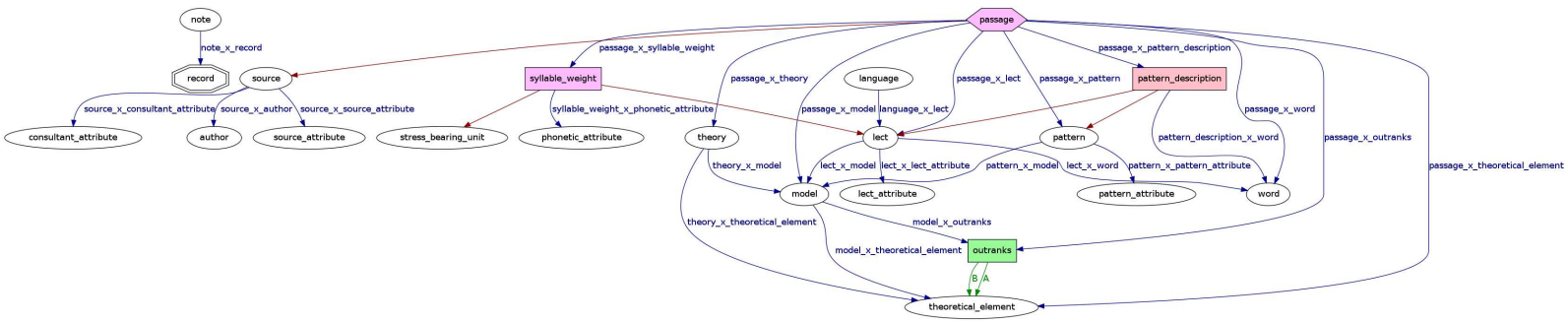
- Quick and easy searching
- Customizable detailed searching
- Export search results

4. Replicability

- It is important that research conducted with ST2 be replicable.
- The ST2 database will periodically be archived and made freely available through an agreement with the Linguistic Data Consortium at the University of Pennsylvania.
- It is recommended that researchers using ST2 for their own research projects use these archived snapshots so that other researchers can replicate results using the identical information source.

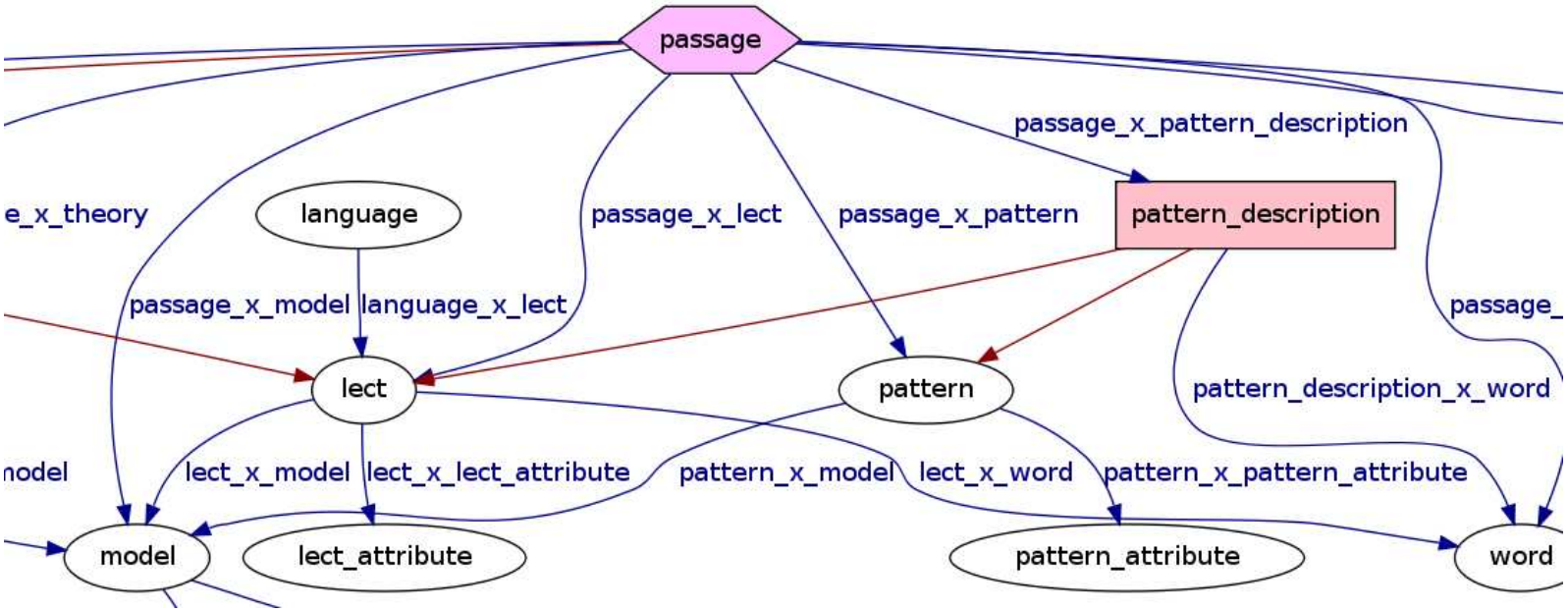
5. Flexibility and 6. Extensibility

Information is organized in a MySQL relational database.



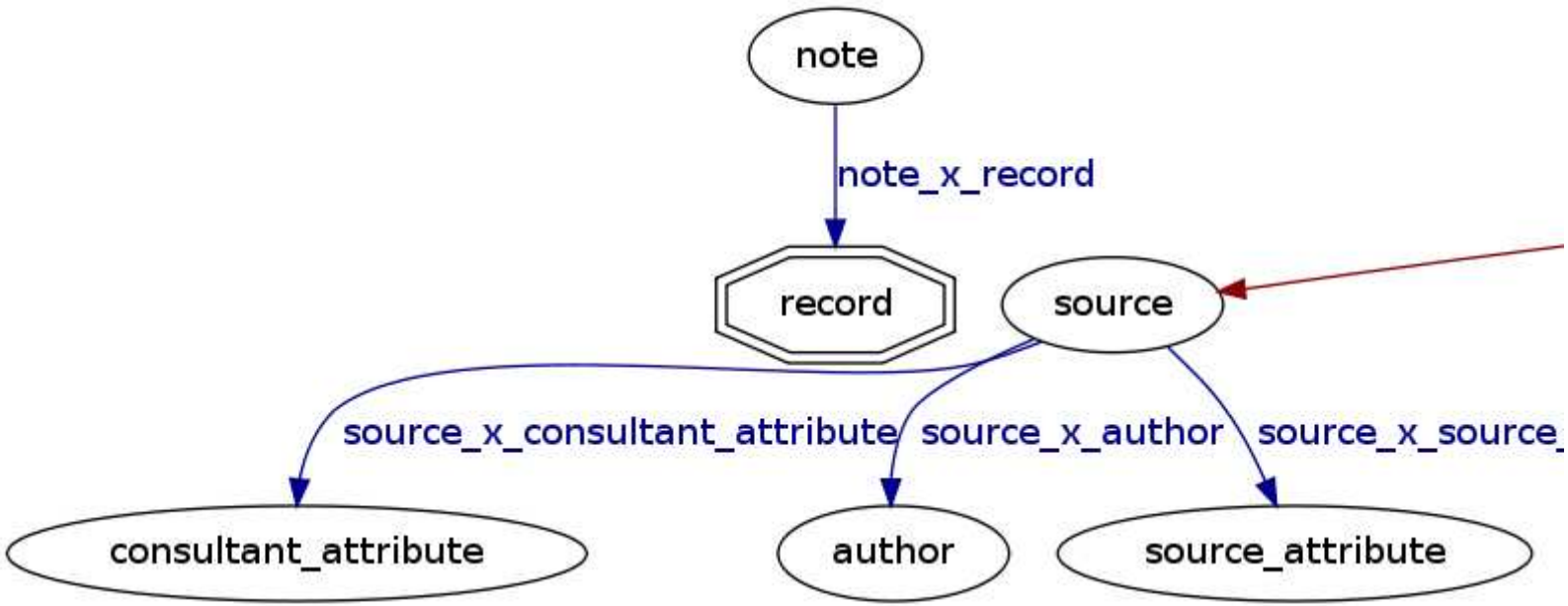
5. Flexibility and 6. Extensibility

1. It distinguishes ‘languages’ as sociopolitical constructs from ‘lects’ as targets of linguistic inquiry.



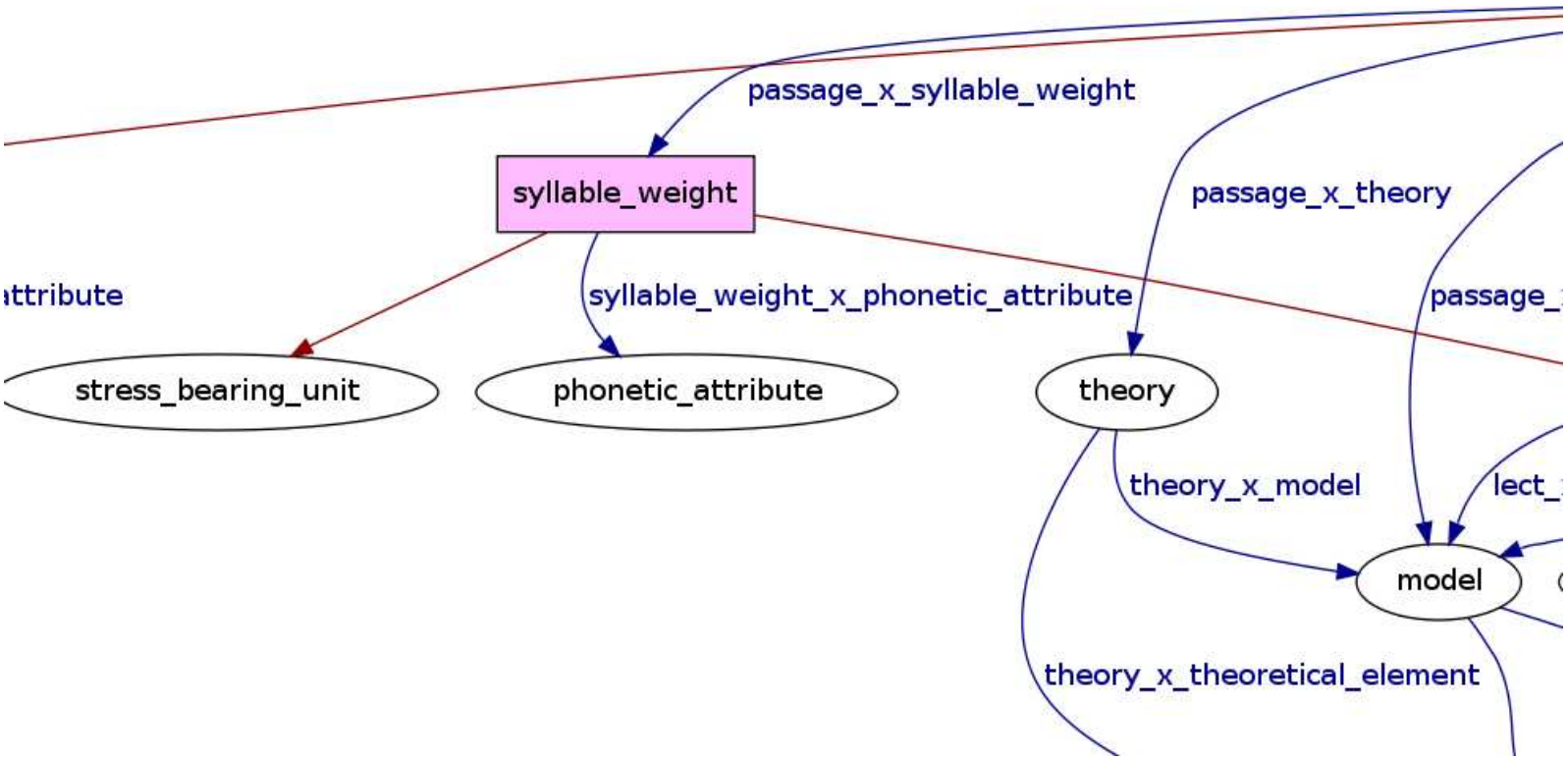
5. Flexibility and 6. Extensibility

2. It allows attributes of sources and consultants which can delimit the scope of studies, if desired.



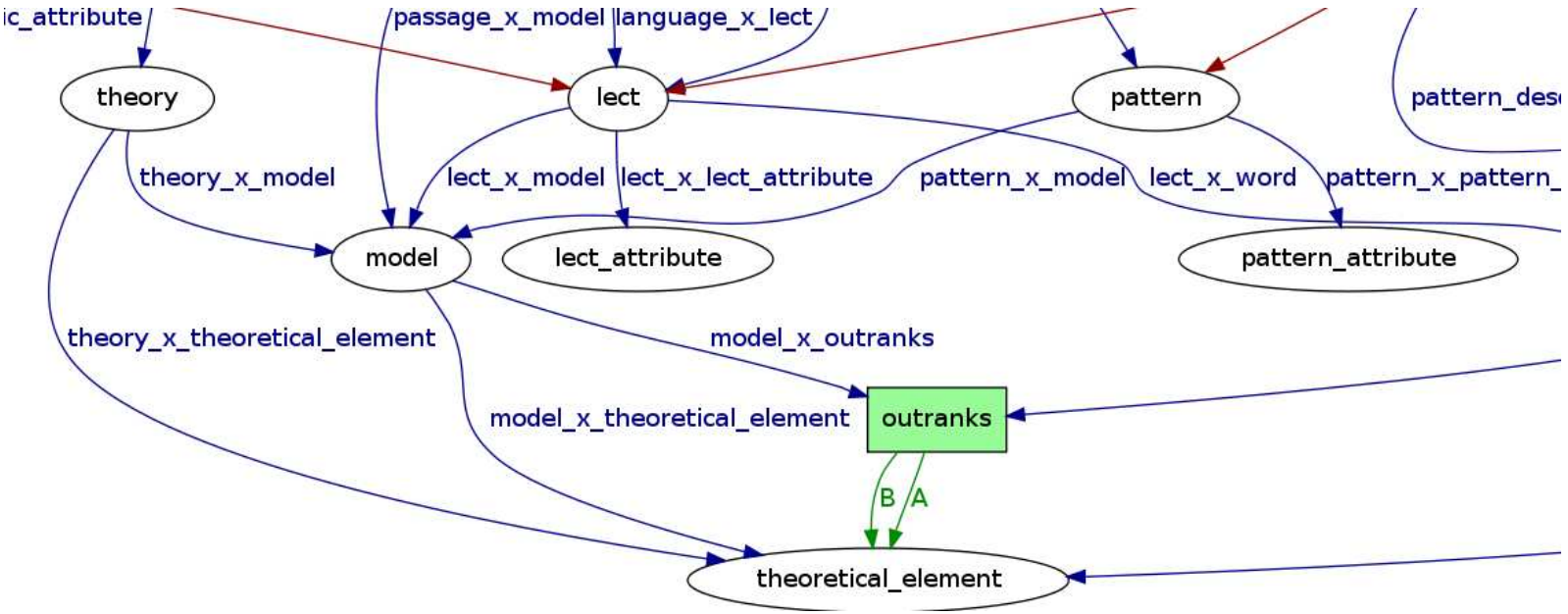
5. Flexibility and 6. Extensibility

3. It allows attributes which describe syllabic and phonetic information.



5. Flexibility and 6. Extensibility

4. ST2 distinguishes 'theories' from 'analyses' (models), allowing new theories and analyses to be added.



Current efforts

- Provide documentation for the web interface and the database
- Clean up the well-studied "tough" cases like English, Dutch, etc.
- Include missing information and correct errors
- Correct errors in the code (debugging)
- Archiving the first version of ST2 with the Linguistic Data Consortium expected next month.

Ongoing and future efforts

- Obtain feedback on data and design so we can continue to develop an ever more useful "Encyclopedia of Types"
- Add new analyses
- Addition of new data on lects and languages (over 100 new lects currently being added)
- Expand ST2 to include pitch accent languages, beginning with the many pitch accent systems in Japanese. (See poster by Jardine and Payne.)

Please Use and Give Feedback!

st2.ullet.net

Feedback can be emailed to **stresstyp2@gmail.com**

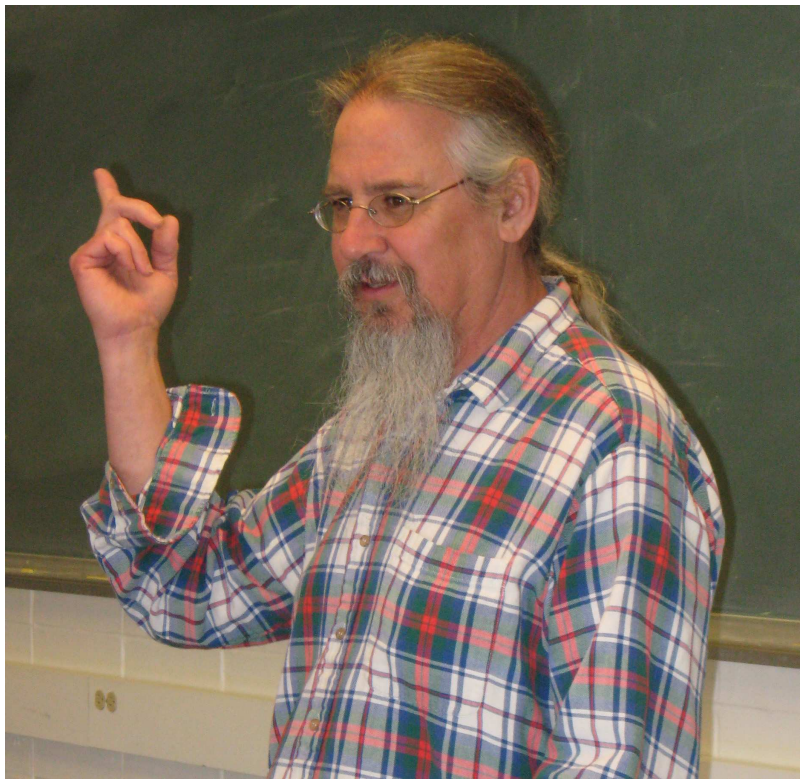
Outline

So what kinds of stress patterns are there?

- Developing hypotheses regarding universals
- Identifying the nature of the variation

Here we will investigate the **computational nature** of the *dominant* stress, rhythm and accent patterns in languages.

Collaborators



Jim

- Prof. Jim Rogers (Earlham College)
- Margaret Cho (Earlham College, BA exp. 2013)
- Sean Wibel (U. Washington, MA exp. 2015)

Modeling stress patterns with stringsets

Example

Penultimate Stress

Primary stress falls on the penultimate syllable and there is no secondary stress.

$$\acute{\sigma}$$
$$\acute{\sigma}\sigma$$
$$\sigma\acute{\sigma}\sigma$$
$$\sigma\sigma\acute{\sigma}\sigma$$
$$\sigma\sigma\sigma\acute{\sigma}\sigma$$
$$\sigma\sigma\sigma\sigma\acute{\sigma}\sigma$$
$$\ldots$$

Modeling stress patterns with stringsets

Example

Leftmost Heavy Otherwise Rightmost (LHOR)

The leftmost heavy syllable is stressed. If there are no heavy syllables, the rightmost syllable is stressed. (E.g. Kwakiutl)

Ĥ	Ĺ	Ĥ L	Ĥ H	L Ĥ
L Ĺ	Ĥ L L	Ĥ L H	Ĥ H L	Ĥ H H
L Ĥ L	L Ĥ H	L L Ĺ	L L Ĥ	L Ĥ L L
L Ĥ L H	Ĥ L L L	Ĥ L L H	Ĥ H L L	Ĥ H L H
L Ĥ H L	L Ĥ H H	Ĥ L H L	Ĥ L H H	Ĥ H H L
Ĥ H H H	L L Ĥ L	L L Ĥ H	L L L Ĺ	L L L Ĥ
...				

Linguistic generalizations describe infinite sets

Linguistic analysis describes these stringsets

1. Every linguistic analysis that generates penultimate stress does so regardless of the length of the word.
2. Likewise, every analysis that generates LHOR does so regardless of the length of the word.

Linguistic generalizations describe infinite sets

Linguistic analysis describes these stringsets

1. Every linguistic analysis that generates penultimate stress does so regardless of the length of the word.
2. Likewise, every analysis that generates LHOR does so regardless of the length of the word.

Also, the infinite set of strings is the point of contact between different analyses that describe the **same** generalization.

Complexity

What are the **properties** of these string sets?

- There is a sense that "LHOR" is a more complex stress pattern than "Penultimate" stress.
- How can we operationalize this insight?

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

Inventories

We can measure the size of the phonemic inventory. It's finite.
Larger inventories are more complex.
(Maddieson 1984, 1992, et seq. . . . Atkinson 2011)

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

But what about sets of strings?

The string sets are of *infinite* size so counting doesn't help!

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

SPE grammars

We can measure the size of a SPE-style grammar by measuring the size of each rule (feature counting). They're finite. Larger grammars are more complex. (Chomsky and Halle 1968)

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

Principles and Parameters

Count the number of parameters needed to be set.

- For example in some metrical theories, QI stress patterns require fewer parameters to be set than QS patterns because QS patterns need to set parameters for which syllables count as heavy, etc.

How can we compare the complexity of different patterns?

One answer: Use size as a proxy for complexity.

Optimality Theory

In OT, phonologies only differ in their ranking. So all are of equal size.

- Counting the number of "active" constraints may be one way to go, but even understanding the effects of simple constraints interacting can be complicated and difficult.
- Perhaps the most concrete approach in this area is T-orders (Antilla 2008)

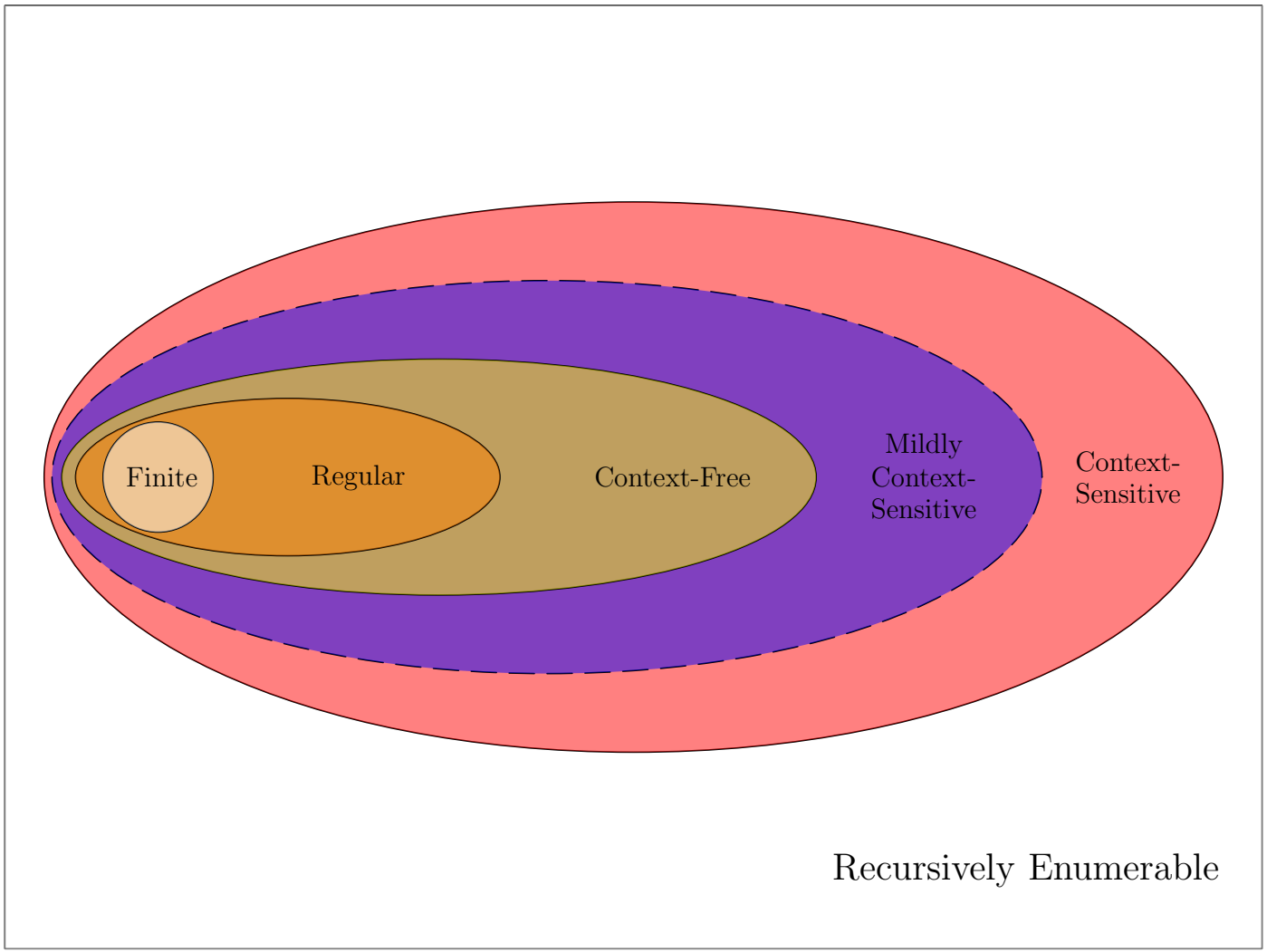
How can we compare the complexity of different patterns?

Computational complexity.

There exist independently-motivated, converging **mathematical criteria** for ordering the complexity of these infinite objects.

- These ideas have been around since the early 1970s (McNaughton and Papert 1971), but were not applied to phonology (until recently).
- These criteria have been argued to be important cognitively (Rogers and Pullum 2011, Rogers et al. 2013, Heinz and Idsardi 2013).
- These criteria are *independent* of any particular mechanism or theory.

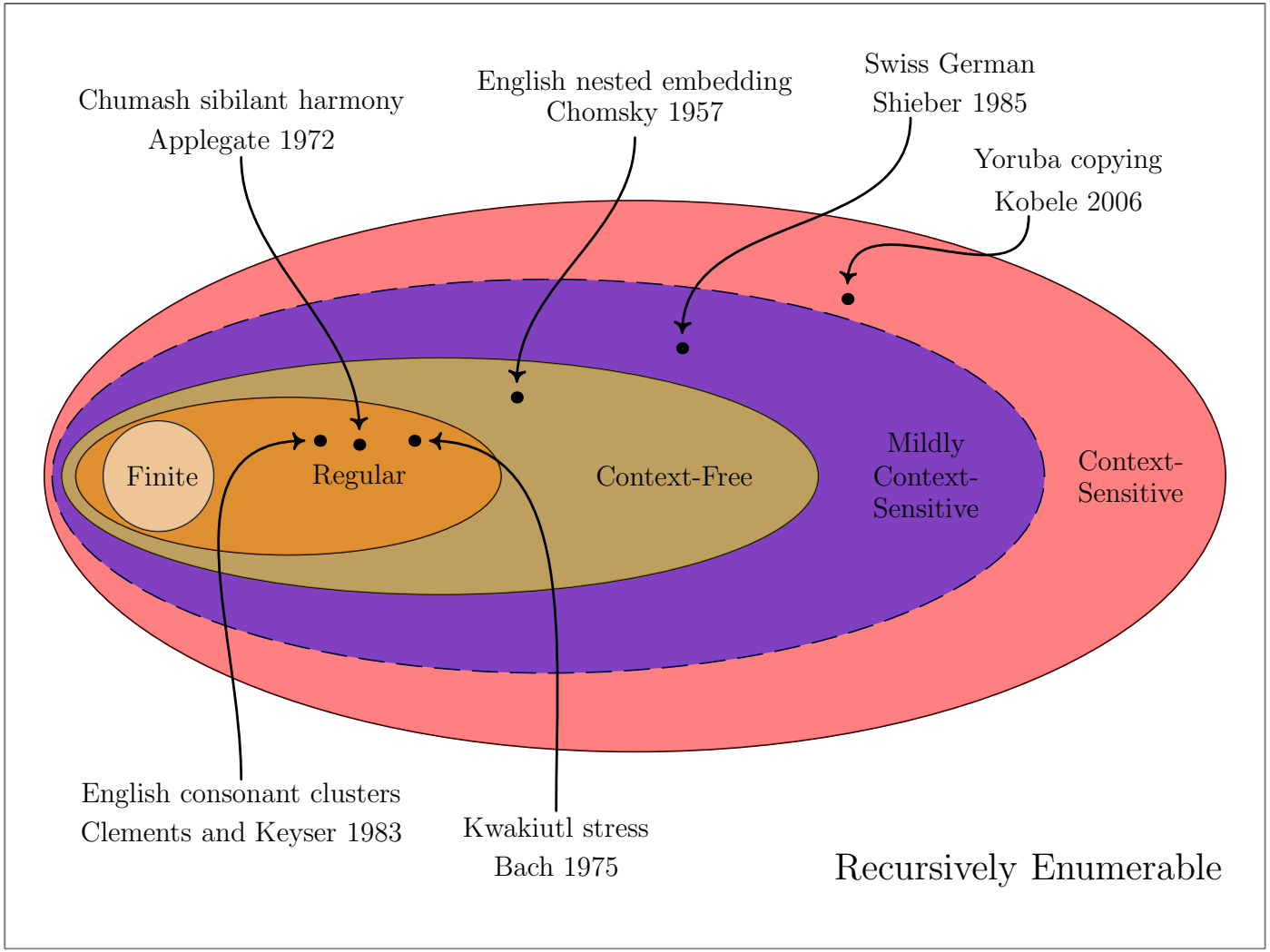
Classifying Sets of Strings



computably enumerable
|
context-sensitive
|
mildly context-sensitive
|
context-free
|
regular
|
finite

Figure: The Chomsky hierarchy

Classifying Sets of Strings



computably enumerable

|

context-sensitive

|

mildly context-sensitive

|

context-free

|

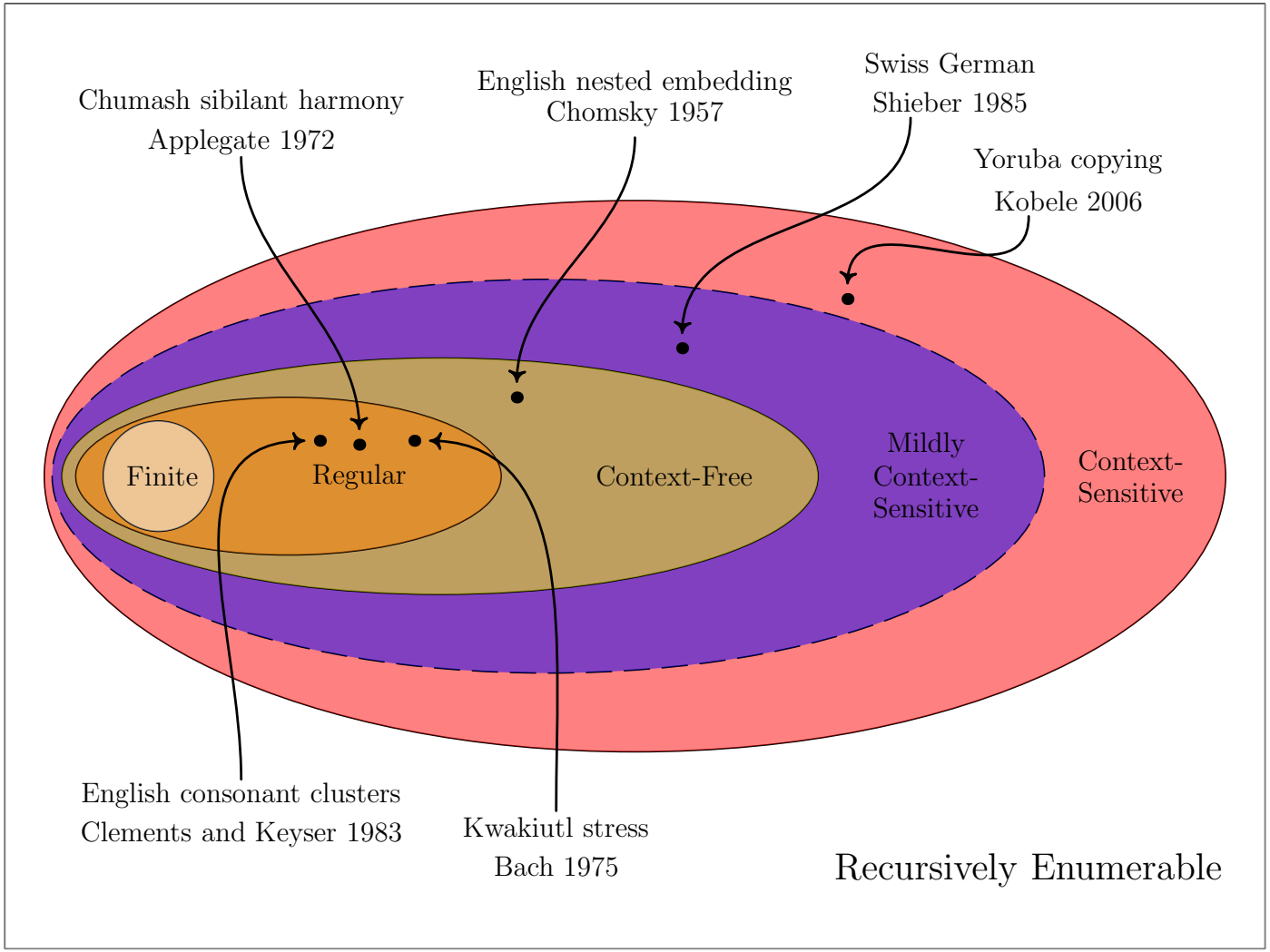
regular

|

finite

Figure: Natural language patterns in the hierarchy.

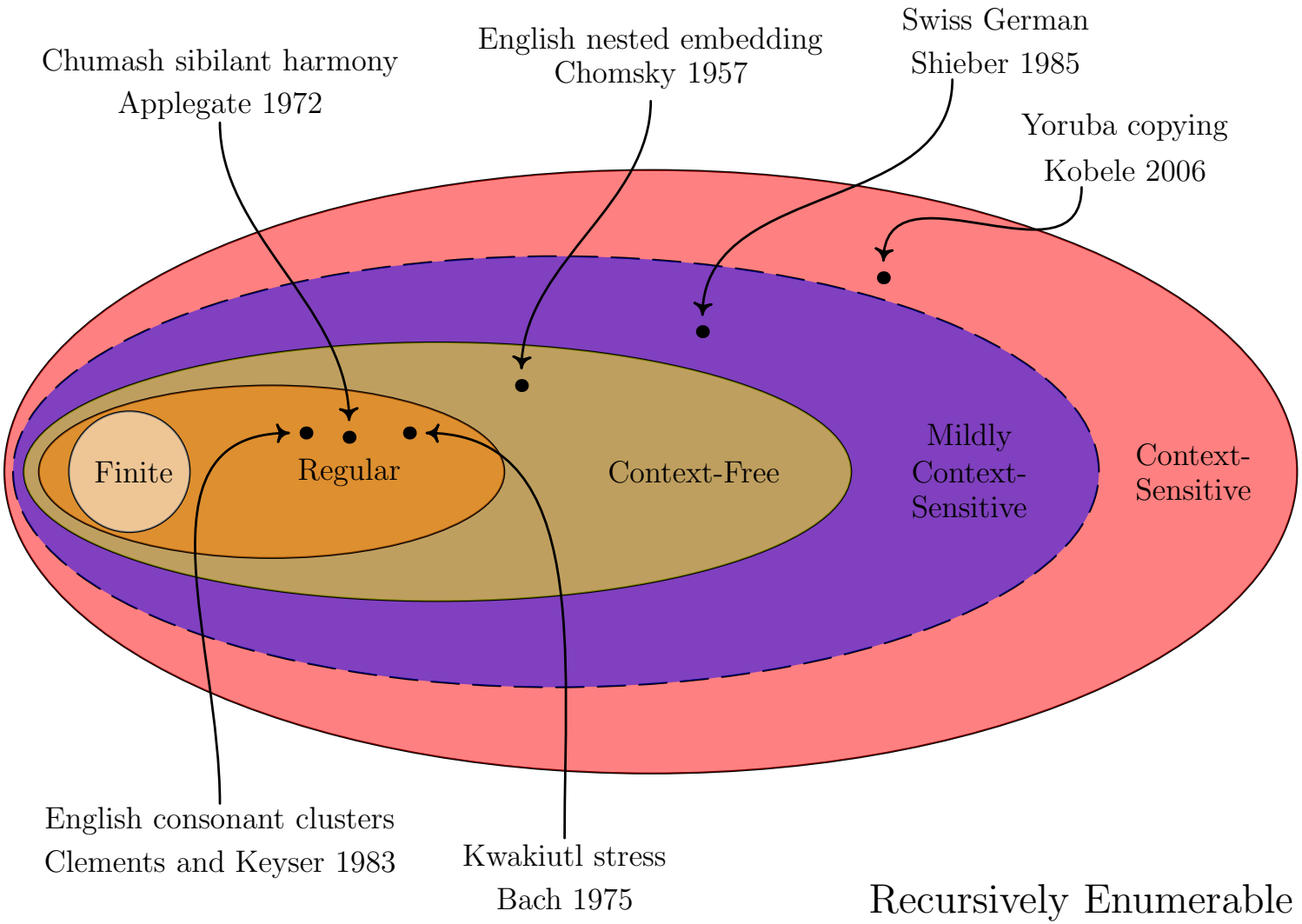
Classifying Sets of Strings



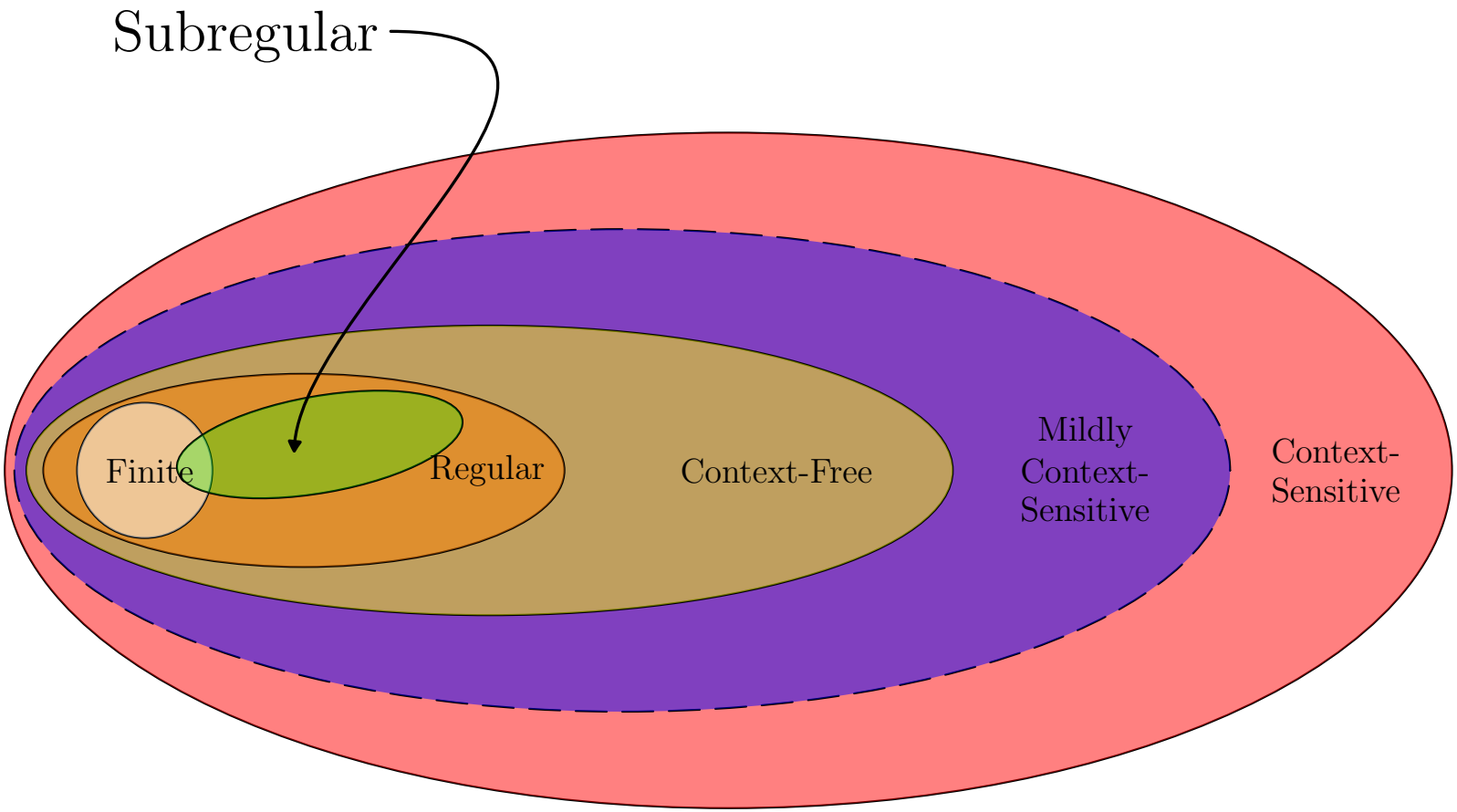
Stress patterns are **regular** (Heinz 2007, 2009).

Figure: Natural language patterns in the hierarchy.

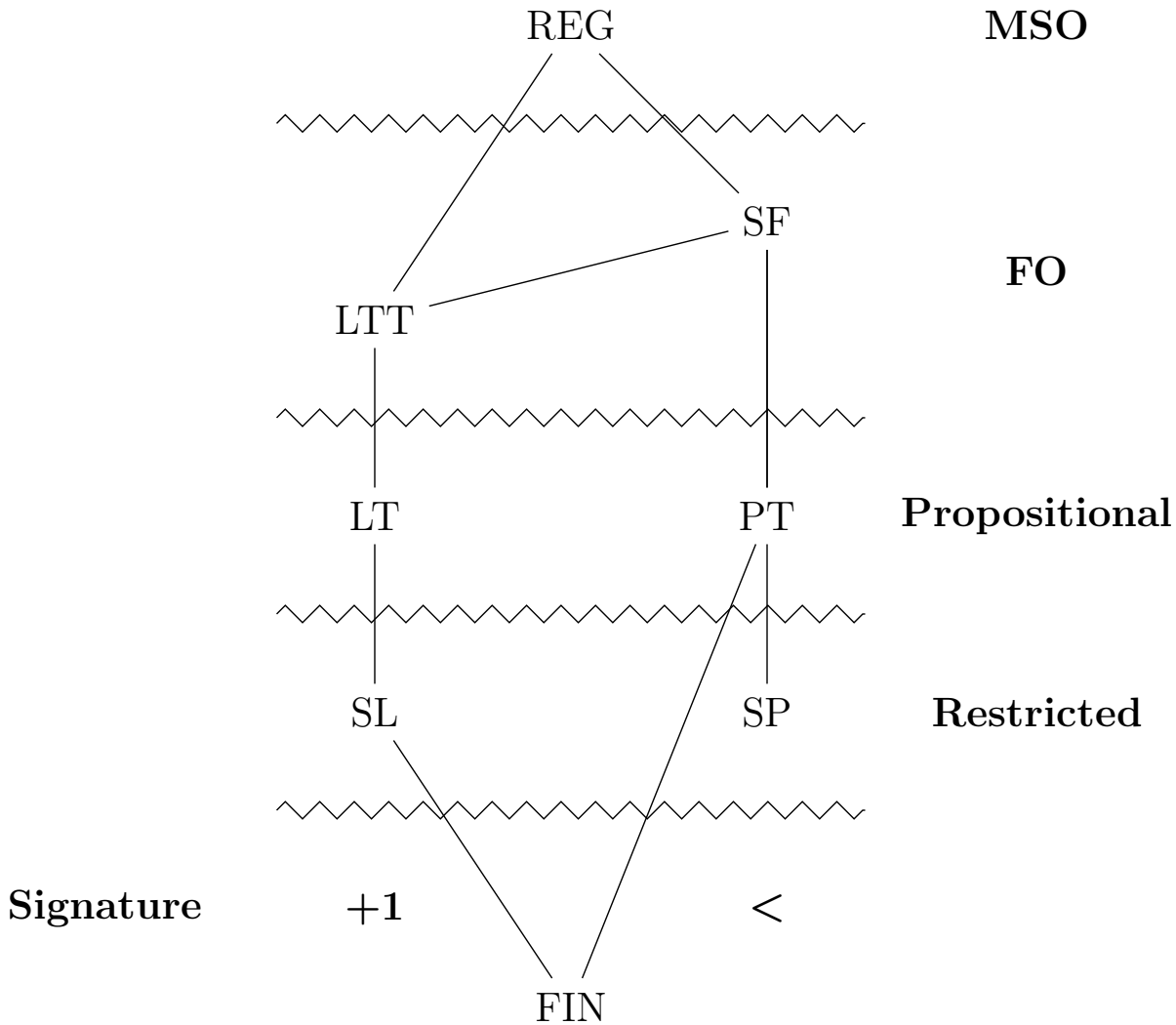
"Being regular" is a start, but it is not sufficient to make the distinctions we want.



"Being regular" is a start, but it is not sufficient to make the distinctions we want.



Encyclopedia of Categories: Sub-regular Stringsets



(McNaughton and Papert 1971, Rogers and Pullum 2011, Rogers et al. 2010, Rogers et al. 2013)

Logical Signatures

The Local Branch (+1)

- (+1) means "successor"
- Literals refer to substrings (contiguous sequences of sounds)

  ex. *σ́σ*, abc

The Piecewise Branch

- (<) means "precedes"
- Literals refer to subsequences (potentially *dis*contiguous sequences of sounds)

  ex. *σ́* . . . *σ́*, a. . . b . . . c

SL and SP: Restricted Logic

Finitely many conjunctions of negative literals define stringsets.

Strictly Local (+1)

ex. $\neg\sigma\sigma\# \wedge \neg\acute{\sigma}\# \wedge \ldots$
Don't have $\sigma\sigma\#$ and don't have $\acute{\sigma}\#$, ...

Strictly Piecewise ($<$)

ex. $\neg\acute{\sigma} \ldots \acute{\sigma} \wedge \ldots$
Don't have $\acute{\sigma} \ldots \acute{\sigma}$ and ...
Don't have two or more primary stressed syllables (Culminativity).

LT and PT: Propositional Logic

Well-formed statements of propositional logic with the literals define stringsets.

Locally Testable (+1)

ex. $\acute{\sigma}$
There is a primary stressed syllable.
Have at least one primary stress (Obligatoriness).

Piecewise Testable (<)

ex. $\mathrm{s} \ldots \mathrm{s} \Rightarrow \int \ldots \int$
IF a word has a $\acute{\sigma} \ldots \sigma$ subsequence THEN it must also have $\acute{\sigma} \ldots \grave{\sigma}$ subsequence.

LTT and NonCounting (SF): First Order Logic

Well-formed statements of first-order logic with the literals define stringsets. (First order is propositional logic with $\forall, \exists$ quantification over individuals.)

Locally Threshold Testable (+1)

ex. $\exists(x, y, z)[x = \grave{\sigma} \wedge y = \grave{\sigma} \wedge z = \grave{\sigma} \wedge x \neq y \neq z]$
Words must have three secondary stressed syllables.

Noncounting ($<$)

ex. $(\forall x)\big[x = \grave{\sigma} \rightarrow (\exists y)[y = \sigma \wedge y < x]\big]$
If a word has $\grave{\sigma}$ then the $\grave{\sigma}$ must be preceded somewhere by a σ.

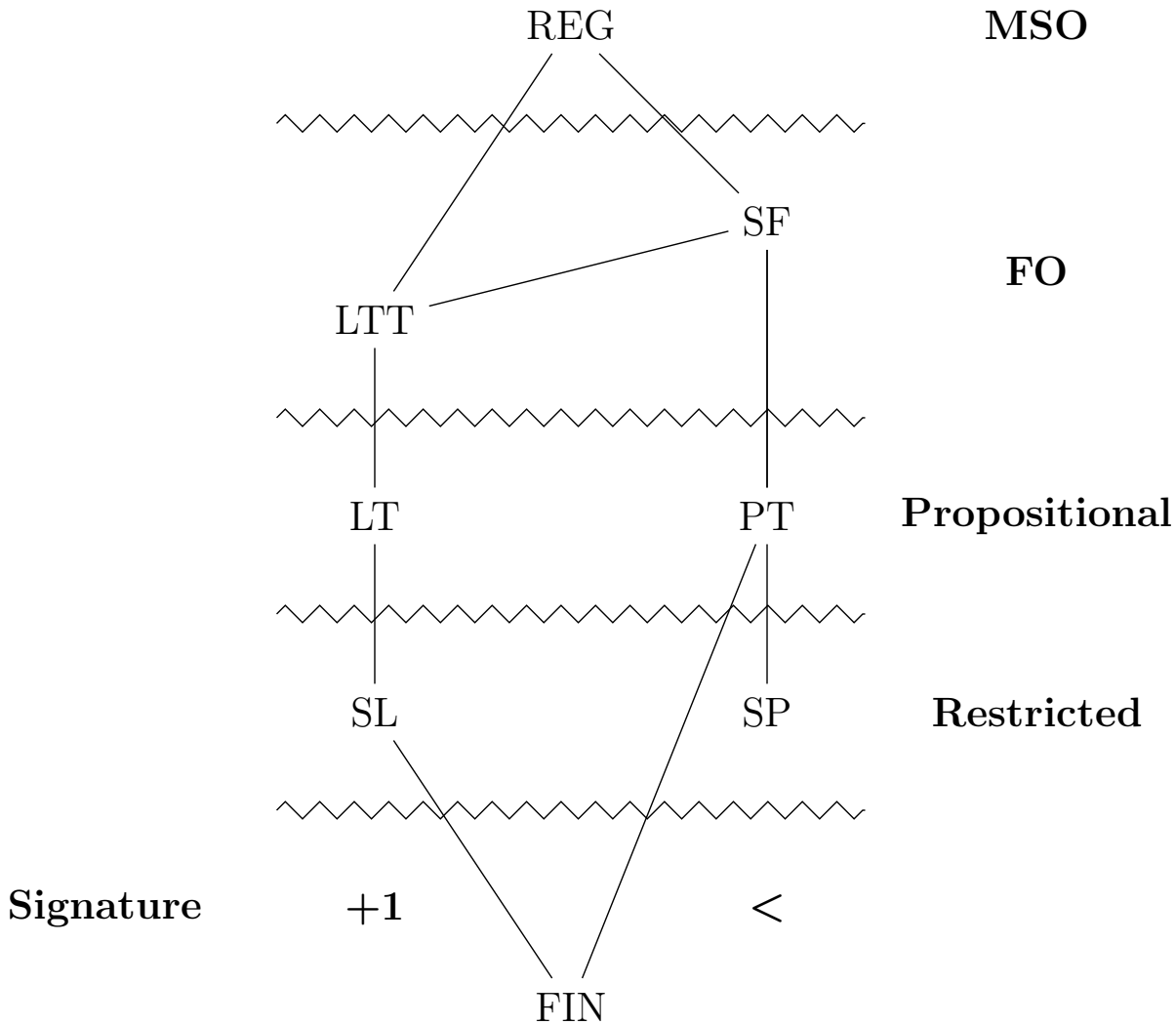
Regular: Monadic Second Order Logic

Well-formed statements of monadic second-order logic with literals from either signature $(+1)$ or $(<)$ define stringsets. (Monadic Second Order is propositional logic with $\forall, \exists$ quantification over *sets* of individuals.)

Regular, either $(+1)$ or $(<)$

ex. Words must have an even number of secondary-stressed syllables.

Encyclopedia of Categories: Subregular Stringsets



(McNaughton and Papert 1971, Rogers and Pullum 2011, Rogers et al. 2010, Rogers et al. 2013)

Typology of (dominant) Stress Patterns

Of the 109 distinct stress patterns studied in Heinz 2009:

- 9 are SL_2. (Initial Stress is here.)
- 44 are SL_3. (Penultimate Stress is here.)
- 24 are SL_4.
- 3 are SL_5. (Asheninca, Bhojpuri, Hindi (Fairbanks))
- 1 is SL_6. (Icua Tupi)
- 28 are not SL_k for any k! These are the *unbounded* patterns like LHOR.

Edlefsen et al. 2009, Rogers et al. 2013, Heinz to appear, Wibel et al. in prep

So how complex are the 28 unbounded patterns?

- The LHOR stringset is properly Noncounting (First Order with $<$)...
- But LHOR reduces to SP_2 *modulo* Obligatoriness (= at least one primary stress).
- In other words, LHOR can be described more simply as the intersection of a stringset which is LT (Obligatoriness) with a stringset which is SP.

(Heinz, to appear)

Factoring the stringsets

- This analysis **factors** complex stringsets into simpler pieces.
- Thus, the complexity of a stringset is given by the complexity of its most complex factor.
- 26 of the 28 remaining patterns are either SP+LT or SL+PT.

(Rogers et al. 2013)

The last two

- The 2 remaining patterns are Cairene Arabic and Creek. They are Counting (Graf 2010). But this result is predicated on whether the secondary stresses are perceptible or not (it's unclear). If they are, then the complexity of these reduces to SL.

Summarizing

Results from the model-theoertic approach

1. With but a few exceptions meriting further attention, the stress patterns in the world's languages belong to either SL, SL+PT, or SP+LT.
2. This result is important for learnability at least in principle provided an upper bound on the length of the (sub)sequence is established.
3. The factorization is yielding about 18 distinct types of stringsets, which we call **primitive constraints**.

Conclusions

1. StressTyp2 presents an encyclopedia of types of stress, accent, and rhythmic patterns in the world's languages.
2. Computer science (model theory) provides an encyclopedia of categories **independent** of any grammatical formalism.
3. From this perspective, there are **restrictive, universal properties** of stress patterns: With only a couple controversial counterexamples, they all can be defined as propositional with $(+1, <)$ signatures.
4. The variation can be limited even further: there appear to be fewer than 20 primitive constraint types.

Conclusions

1. StressTyp2 presents an encyclopedia of types of stress, accent, and rhythmic patterns in the world's languages.
2. Computer science (model theory) provides an encyclopedia of categories **independent** of any grammatical formalism.
3. From this perspective, there are **restrictive, universal properties** of stress patterns: With only a couple controversial counterexamples, they all can be defined as propositional with $(+1, <)$ signatures.
4. The variation can be limited even further: there appear to be fewer than 20 primitive constraint types.

Conclusions

1. StressTyp2 presents an encyclopedia of types of stress, accent, and rhythmic patterns in the world's languages.
2. Computer science (model theory) provides an encyclopedia of categories **independent** of any grammatical formalism.
3. From this perspective, there are **restrictive, universal properties** of stress patterns: With only a couple controversial counterexamples, they all can be defined as propositional with $(+1, <)$ signatures.
4. The variation can be limited even further: there appear to be fewer than 20 primitive constraint types.

Conclusions

1. StressTyp2 presents an encyclopedia of types of stress, accent, and rhythmic patterns in the world's languages.
2. Computer science (model theory) provides an encyclopedia of categories **independent** of any grammatical formalism.
3. From this perspective, there are **restrictive, universal properties** of stress patterns: With only a couple controversial counterexamples, they all can be defined as propositional with $(+1, <)$ signatures.
4. The variation can be limited even further: there appear to be fewer than 20 primitive constraint types.

Thank you for listening!