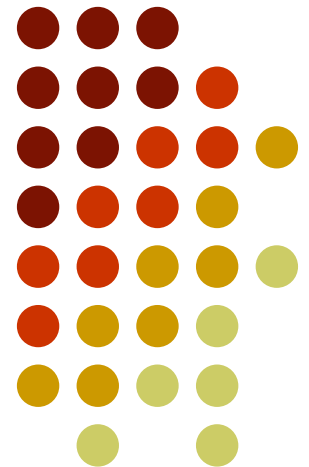


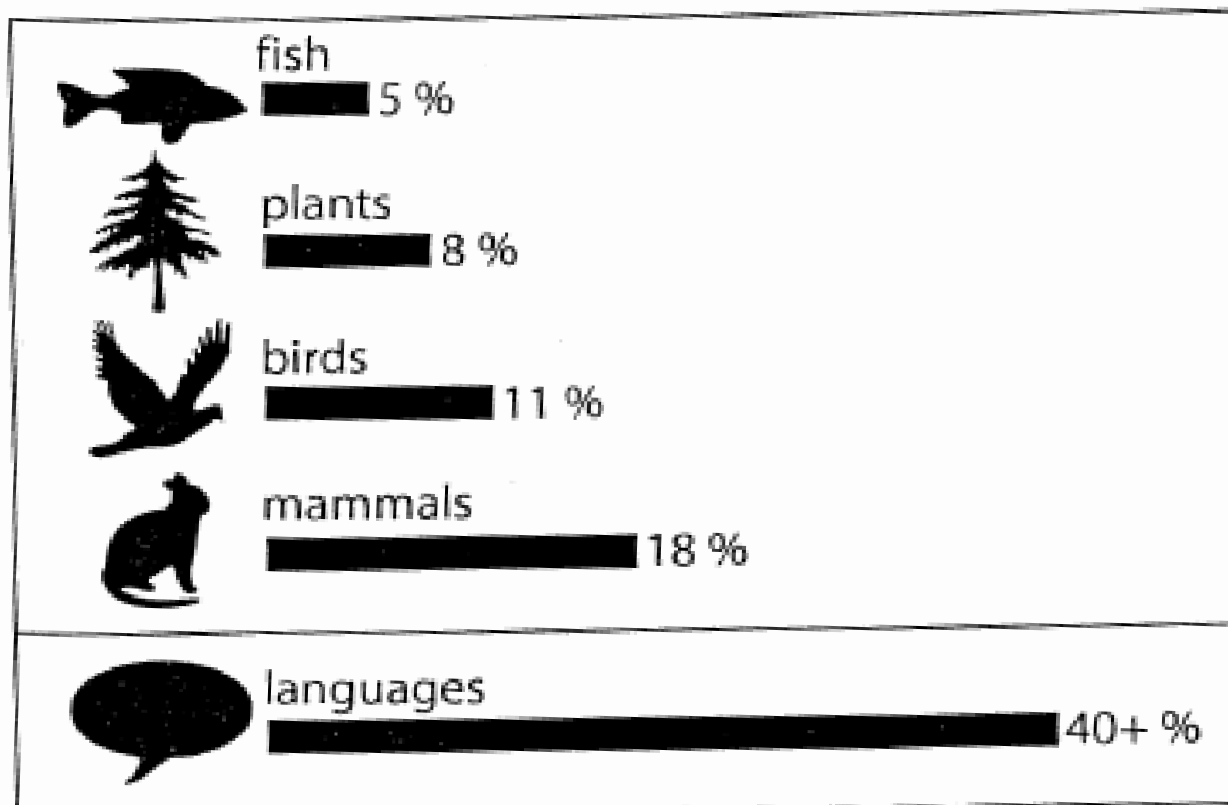
Developing a corpus for a morphologically rich, endangered language



Deryle Lonsdale / Dawn Bates
BYU / Arizona State University
lonz@byu.edu / dawn.bates@asu.edu



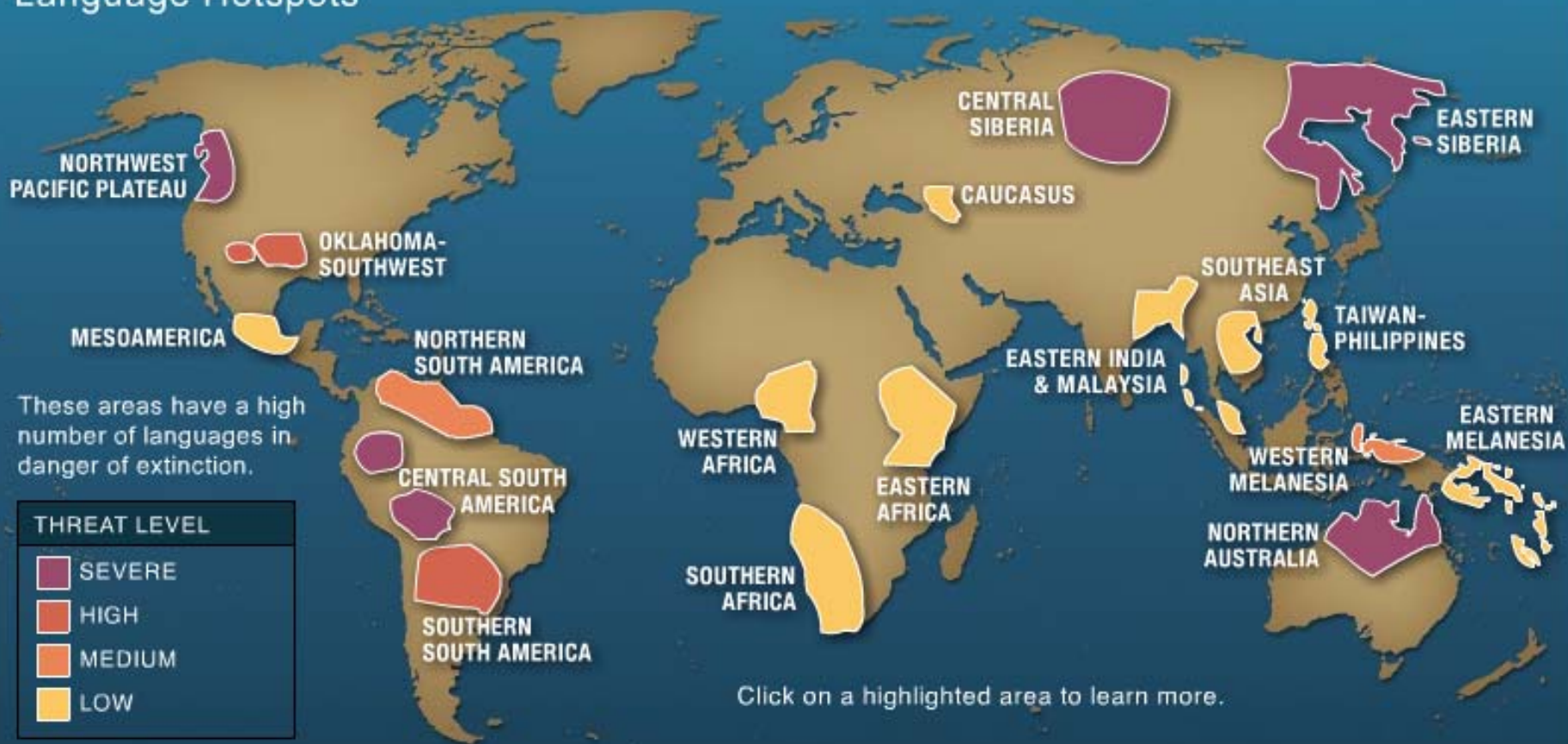
Language endangerment



Language hot-zones

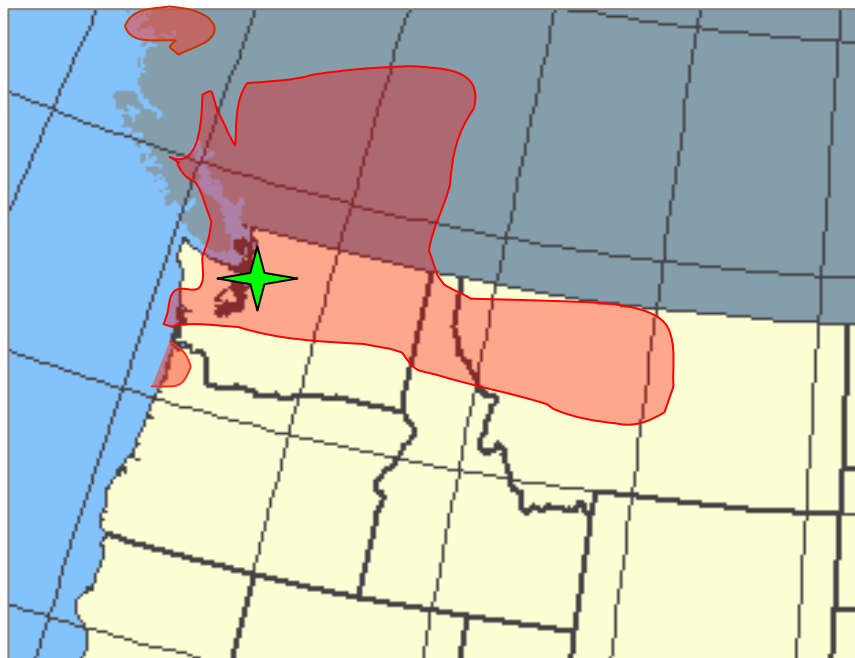


Language Hotspots



Source: Living Tongues Institute for Endangered Languages

The Salish language family



✦: Lushootseed

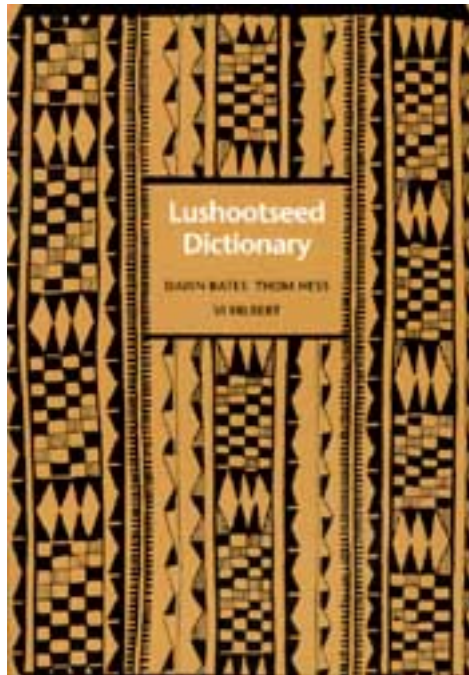
- 23 languages
- Coastal & Interior
- 5 major subfamilies
- Many moribund
- Most un-IE (Coastal)
- Intensive corpus work (speech, text, video, archives)
- Ongoing active analysis

Lushootseed (dx^wləšucid)



- Also called Puget Salish
 - Two dialect groups (north, south)
 - Highly endangered, active work
 - First linguistic grammar is still in preparation
 - Spoken by tens of thousands of speakers 150 years ago
 - Today only a handful of L1 speakers, all elderly
 - Active revitalization effort, many L2, heritage speakers
 - Maybe 500K written words extant
- Lushootseed dialects:
 - Duwamish
 - Kikiallus
 - Muckleshoot
 - Nisqually
 - Nuwaha
 - Puyallup
 - Sahehwamish
 - Skagit
 - Sauk-Suiattle
 - Skykomish
 - Snohomish
 - Snoqualamie
 - Squaxin
 - Steilacoom
 - Stillaguamish
 - Suquamish
 - Swinomish

Lushootseed dictionary (LD)



- Resource intended for students, researchers
- Update to previous version
- Recent work on digitizing, annotating contents
- Crucial tool for research in this talk



Our work

- Collect and annotate a corpus of Lushootseed sentences
 - 1) Find sentences from published sources
 - 2) Perform a morphological parse on the input words
 - 3) Perform a syntactic parse on the morphologically parsed words
 - 4) Convert the annotations to a database format
 - 5) Query the database for various types of information

Rehabilitating the lexical data



- From LEXWARE to TEI (XML)
- From outdated character set representation to Unicode
- Searchable, browsable digital version
- Lexical content, usage examples used in corpus work



1) The source sentences

- Randomly chosen from several sources
 - Conversational turns from published narratives
 - Sample sentences from pedagogical grammars
 - Example usage sentences from LD
 - Even historical manuscripts where possible
- Typed or scanned into Romanized form
- No systematically balanced coverage



2) Parsing the words

- Each sentence is tokenized and each word sent through a finite-state parser
 - Rules to describe morphophonological changes
 - Lexicons to define roots, morpheme types
 - Root coverage via Bates/Hess/Hilbert
- The result is a morphological parse for each incoming sentence.



Salish morphology

- Notoriously famous: polysynthetic
- Roots: simple; mostly 1, 2 syllables
- Extensive inflection, derivation
- Intricate reduplication patterns, vowel harmony, valency modifications
- Lexical suffixes: special class of bound morphemes that have lexical referents



Sample rule, table, FSA

;;; Optional syncope rule
 ;;; Note: free variation
 ;;; L: Lu+ad+s+pastEd
 ;;; S: L00ad0s0pastEd

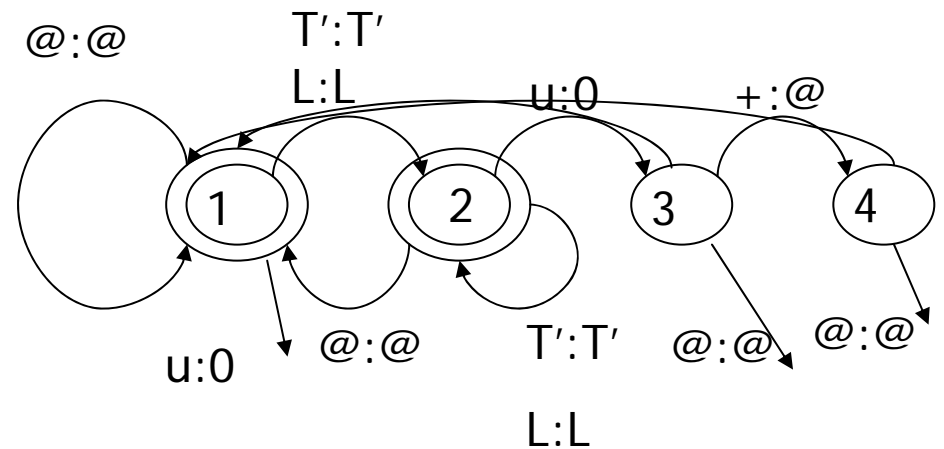
RULE

"u:0 => [L|T'] __ +:@ VW" 4 6

u:0
VW:VW

u:0

	u	L	+	VW	@	T'
0	L	@	VW	@	T'	
1:	0	2	1	1	1	2
2:	3	2	1	1	1	2
3.	1	0	4	0	0	0
4.	1	0	0	1	0	0



Sample parses (Romanized)



PC-KIMMO>recognize gWEdsutudZildubut

gWE+ d+ s+ ?u+ ^tudZil +du +b +ut ←

Dub+ my+ Nmz+ Perf+ ^bend_over +OOC +Midd +Rfx

PC-KIMMO>recognize adsukWaxWdubs

ad+ s+ ?u+ ^kWaxW +du +b +s ←

Your+ Nmz+ Perf+ ^help +OOC +Midd +his/hers

Morphology of the sentence



1) Romanize the sentence.

tuLildExW kWi ?aciLtalbixW ?E kWi lEpEskWi? .

2) Morphologically decompose the sentence.

tu+ Lil +d +ExW kWi ?aciLtal =bixW ?E kWi lEpEskWi? .



3) Parsing the syntax

- Describe the relationships between the words and the morphemes in a sentence.
- Many possible approaches
- I chose the Link Grammar parser (Sleator & Temperley 1993).
- Doesn't build trees, but diagrams pair-wise associations between elements in the sentence
- Most flexible framework for addressing Lushootseed constructions



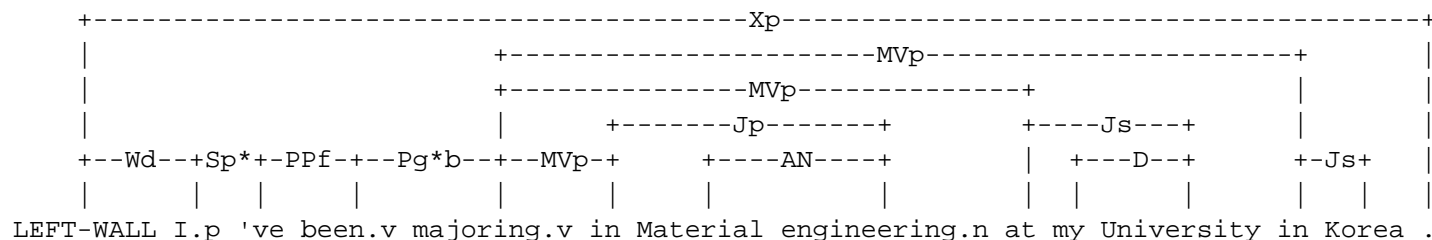
The Link Grammar parser

- A free engine for performing syntactic parsing
- Fast, efficient, widely used
- Needs a lexicon
 - Used the same information as for the morphology engine lexicon
- Needs a specification of link types and how they combine
 - Has to be developed for each language; not for the faint-hearted!

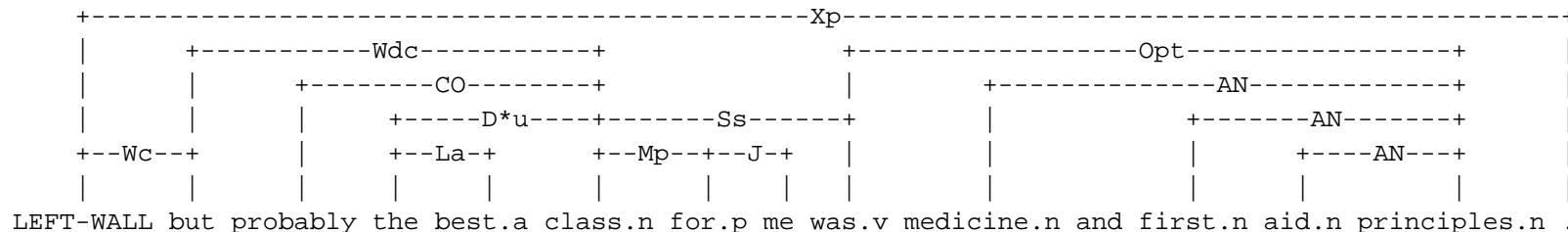
LG example parses (English)



Linkage 1, cost vector = (UNUSED=0 DIS=2 AND=0 LEN=23)



Linkage 1, cost vector = (UNUSED=0 DIS=2 AND=0 LEN=27)



Parsing Lushootseed with LG



- Two steps
 1. Process the words morphologically (identify morphemes, break apart words)
 2. Build links between words in the sentences (run the LG engine on the incoming sentence)
- Uses
 - Specifications for link types (the grammar)
 - Specifications for word types (the lexicon)

Sample LG Lushootseed parse



```
linkparser> tu+ Lil +d +ExW kWl ?aciLtalbixW ?E kWl lEpEskWi?.  
++++Time (0.13 total) 0.06 seconds  
Found 15 linkages (15 had no P.P. violations)  
Linkage 1, cost vector = (UNUSED=0 DIS=4 AND=0 LEN=19)
```

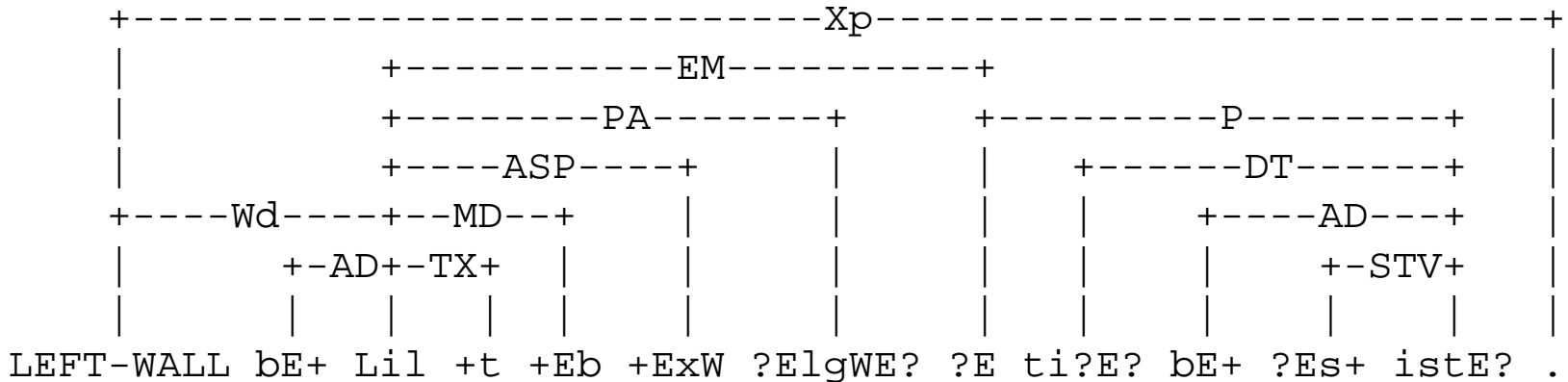
```
+-----Xp-----+  
|               +-----EX-----+  
|               +-----SOS-----+ |  
+----Wd----+--ASP--+ |               +----P----+  
|          +-PT+-TX+ |          +---DT---+ |          +---DT---+  
|          |   |   | |          |   |   | |          |   |   |  
LEFT-WALL tu+ Lil +d +ExW kWl ?aciLtalbixW ?E kWl lEpEskWi? .
```

Press RETURN for the next linkage.



Another parse

```
linkparser> bE+ Lil +t +Eb +ExW ?ElgWE? ?E ti?E? bE+ ?Es+ istE?.
++++Time    0.07 seconds (0.20 total)
Found 2 linkages (2 had no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=2 AND=0 LEN=28)
```





Parsing Lushootseed

- How? Problematic...
 - No integrated comprehensive theoretical account of morphology/syntax and interactions
 - Lexical information necessary
 - What corpus?
- Why? Debatable, but...
 - Insight into structure
 - Reference for linguistic research
 - Potential for language learning applications
 - Statistical and distributional information



Parser development

- Collect as many sentences as possible
- Run them through the morphological and syntactic parsers
- Collect the results for further analysis
- Refine the rules as necessary

- So far: about a thousand Lushootseed sentences parsed



Another parse

```
linkparser> q'ili +t +Eb +ExW ?E ti?E? s+ ?il =aXad ?E ti?E? captain.  
++++Time 0.08 seconds (0.28 total)  
Found 11 linkages (11 had no P.P. violations)  
Linkage 3, cost vector = (UNUSED=0 DIS=6 AND=0 LEN=23)
```

```
+-----Xp-----+  
|           +-----EM-----+           |  
|           +-----ASP-----+ +-----P-----+           |  
|           +---MD---+           |           +---DT---+-----MV---+-----P-----+           |  
+---Wd---+---TX---+           |           |           +NZ+-LX-+           |           +---DT---+           |  
|           |           |           |           |           |           |           |           |           |           |  
LEFT-WALL q'ili +t +Eb +ExW ?E ti?E? s+ ?il =aXad ?E ti?E? captain .
```


Sample link specifications



```
<pref-asp1>: {(PRF- or STV- or PRG-)};
<pref-asp2>: {HAB-} & {DUB-} & {AD-};
<predprefs>: {NZ-} & {<pref-asp1>} & {SX-} & {<pref-asp2>} & {(FUT- or PT-)};

<root-main>: <predprefs> & {DT-} & {LX+} & {BNF+} & {TX+} & {TC+} & {ACH+} & {TC+} & {TX+} &
  {ASP+};
<main-args>: {P-} & {GEN-} & {WH-} & {SOs+} & {MV+};

<root-ditrx>: <predprefs> & {DT-} & {LX+} & {BNF+} & TX+ & {TC+} & {ACH+} & {TC+} & {TX+} &
  {ASP+};
<ditrx-args>: {P-} & {GEN-} & {WH-} & {SOs+} & {EX+} & {SOo+} & {MV+};

<root-middle>: <predprefs> & {DT-} & {LX+} & {BNF+} & {TX+} & {TC+} & MD+ & {ACH+} & {TC+} &
  {TX+} & {ASP+};
<middle-args>: {P-} & {GEN-} & {WH-} & (({PA+} & {EM+}) or ({EM+} & {PA+})) & {MV+};

<pred1>: ((<root-main> & <main-args>) or
  (<root-middle> & <middle-args>) or
  (<root-ditrx> & <ditrx-args>))
  & {Wd-};
```



4) Database conversion

- The link structure for each word pair can be loaded into a database record
 - Import parse results into MS Access
- This allows use of database manipulation techniques
- Querying, or asking about, the contents is a commonly performed task with databases

Microsoft Access - [gramruth2 : Table]

File Edit View Insert Format Records Tools Window Help

Type a question for help

ID	SentNum	LHLex	LHLabel	LHCount	Label	RHCount	RHLabel	RHLex
1842	356	T'u+	HAB	1	<---HAB--->	3	HAB	t'as.r
1843	356	tu+	PT	2	<---PT---->	3	PT	t'as.r
1063	206	huy.a	Wd	1	<---Wd---->	2	Wd	tEJ.r
1068	207	gWEI	Wd	1	<---Wd---->	2	Wd	tEJ.r
406	74	gWEI	Wd	1	<---Wd---->	2	Wd	tEL.r
681	123	gWE+	DUB	1	<---DUB--->	2	DUB	tEL.r
1498	297	gWE+	DUB	1	<---DUB--->	2	DUB	T'EI.r
24	4	?E	P	6	<---P----->	8	P	t'EqxW.r
25	4	s+	NZ	7	<---NZ---->	8	NZ	t'EqxW.r
1305	263	?al	P	10	<---P----->	12	P	t'Es.r
1306	263	kWi	DT	11	<---DT---->	12	DT	t'Es.r
1546	305	bE+	AD	1	<---AD---->	2	AD	tEy.r
1976	378	gWEI	Wd	1	<---Wd---->	2	Wd	ti?E?.d
212	36	IEk'W.r	SOs	5	<---SOs--->	8	SO	ti?E?.p
216	37	?a.r	SOs	1	<---SOs--->	2	SO	ti?E?.p
421	76	IEk'W.r	SOs	2	<---SOs--->	6	SO	ti?E?.p
491	89	dxW?al	P	5	<---P----->	6	P	ti?E?.p
522	94	gWaXW.r	SOs	1	<---SOs--->	3	SO	ti?E?.p
548	97	gWEIub.r	SOs	2	<---SOs--->	3	SO	ti?E?.p
560	99	huy.r	SOs	6	<---SOs--->	8	SO	ti?E?.p
561	99	?E	P	7	<---P----->	8	P	ti?E?.p
1351	271	LixW.r	SOs	1	<---SOs--->	2	SO	ti?E?.p
1420	283	?al	P	3	<---P----->	4	P	ti?E?.p
1430	285	?al	P	3	<---P----->	4	P	ti?E?.p
1437	287	?al	P	3	<---P----->	4	P	ti?E?.p
1786	349	?ab.r	PA	3	<---PA---->	8	PA	ti?E?.p
1846	357	?ay.r	PA	2	<---PA---->	7	PA	ti?E?.p
105	15	?E	P	5	<---P----->	9	P	T'ub.r
106	15	studEq.r	SOs	7	<---SOs--->	9	SO	T'ub.r
108	15	Lu+	FUT	8	<---FUT--->	9	FUT	T'ub.r
1986	381	?Es+	STV	1	<---STV--->	2	STV	T'ub.r
132	18	Jiq'.r	SOs	1	<---SOs--->	9	SO	tudEq
137	18	?E	P	6	<---P----->	9	P	tudEq
138	18	ti?E?.d	NT	7	<---NT---->	9	NT	tudEq

Record: 834 of 2143

Datasheet View NUM





5) Query over the analyses

- Once in database format, the parses can be queried by a user using SQL:
 - Which predicates have both a negative and an aspectual marker?
 - Which sentences have two oblique complements?
 - Find questions with past tense.
 - Which words are the most complex?
- SQL is not English; some technical knowledge is necessary



Sample SQL query

- List sentences with respect to the complexity of their predicates

```
SELECT SentNum,COUNT(SentNum) from gramruth2
WHERE (LHLex LIKE "*+" OR LHLex LIKE "*=")
OR
(RHLex LIKE "+*" OR RHLex LIKE "=*")
GROUP BY SentNum;
```

Sample statistics (tokens)



- # sentences: 500
- # morphemes: 2954
 - suffixes: 623
 - prefixes: 607
 - lexical suffixes: 58
- # words: 1625
- # S's with only monomorphemic words: 43



Sample statistics (links, partial)

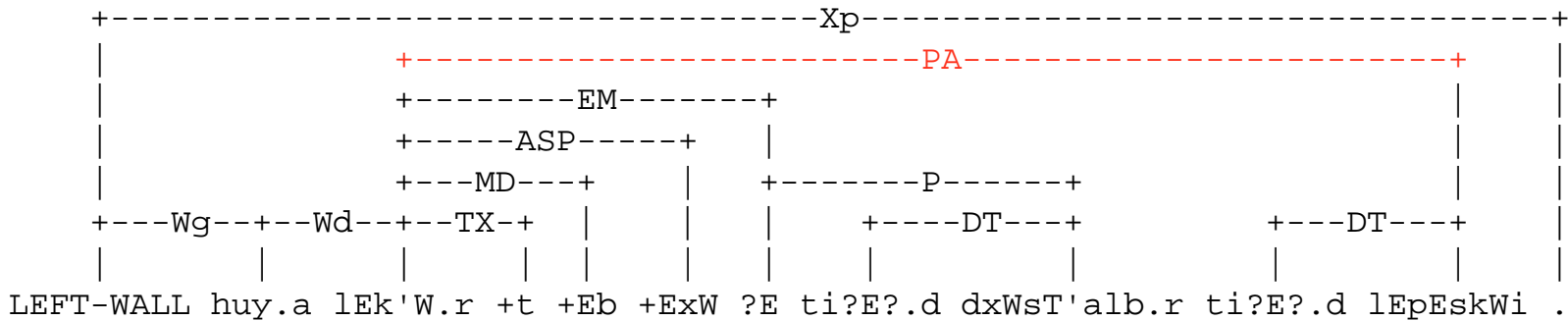
777	punctuation	73	lexical suffix
763	determiner	66	oblique
629	PP	61	habitual
618	subject	53	subordinating
528	PP-object	51	adverbial (predicate)
380	aspectual	45	passive
323	transitive	39	dubitative
228	middle	32	benefactive
205	stative	29	progressive
193	nominalizer	29	causative
188	past	15	object
122	adverbial (sentential)	10	adjective
99	achievement	8	determiner (feminine)
97	perfective	7	partitive
85	possessive	2	reflexive
82	future		



Sample structures

- Longest link

```
linkparser> huy lEk'W +t +Eb +ExW ?E ti?E? dxWsT'alb ti?E? lEpEskWi.
```



- Most complex predicate

```
T'u+ tu+ s+ takW+ +yi +Eb +s
```

Lexical suffixes frequency?



```
10 =bixW
  3 =igWEd
  2 =ELdat
  2 =a?kW
  2 =aCi?
  2 =aXad
  2 =ali
  1 =abac
  1 =al?txW
  1 =alikW
  1 =alus
  1 =aq
  1 =gWas
  1 =gWiL
  1 =gWil
  1 =i
  1 =iC
  1 =qid
  1 =ucid
```



Other sample questions

- Find all sentences where a given word is translated into English the same way.
- Show cases where a passive in English is used to translate a middle in Lushootseed.
- Which is more common: past tense or present tense?
- Find all verbs with the out-of-control suffix.
- List all reduplicated forms by pattern.



Issues

- Ethical: native copyright and language ownership
- Representational: XML markup vs. relational database for canonical representation
- Technological: multi-layer annotation of multimedia (audio, video)
 - TASX, EXMARaLDA, CLaRK, LTXML2, IGT-XML, etc.



Future work

- More complete corpus, better coverage
- Unicode orthography
- Linguistic phenomena (e.g. reduplication)
- Menu-driven queries instead of SQL
- Show how it can be useful for language learning, other applications (e.g. speech recognition)
- Deploy it on the web for others to use



Conclusions

- Preliminary work on an annotated corpus of Lushootseed text
- Combines morphological and syntactic parsing
- Well suited for shallow morphosyntactic parsing
- Potentially useful tool for text analysis, language learning and documentation

Questions?

