Building a Large Multilingual Resource using (Semi-)Automated Methods: Finding, Enriching, Repurposing

William Lewis

Microsoft Research

&

Fei Xia

University of Washington

Main Ideas

- Finding Linguistic Data on the Web
- Extracting and Databasing the Data
- Enriching the Data (e.g., through projections)
- Providing Query Facility over the Data
- Training Tools over the Enriched Data

Linguistic Data on the Web

- Large amount of linguistically analyzed language data making its way to the Web
- Not easy to locate, especially if language data embedded in other resources & documents
  - Search engines may locate resources
  - But results noisy and sometimes difficult to ferret through
  - Made more difficult because of the lack of consistency in encoding and presenting data

Linguistic Data on the Web

- Problems:
  - How to make the wealth of language data on the Web *easily* locatable
  - How to provide a search facility across data and repurpose the data (*interoperate)*
- Solutions:
  - Adapt existing technologies to locate resources (Web pages, documents, etc.)
  - Extract, enrich and index data (by language, family, construction, resource)
  - Expose the data to services (search, tool building, etc.)

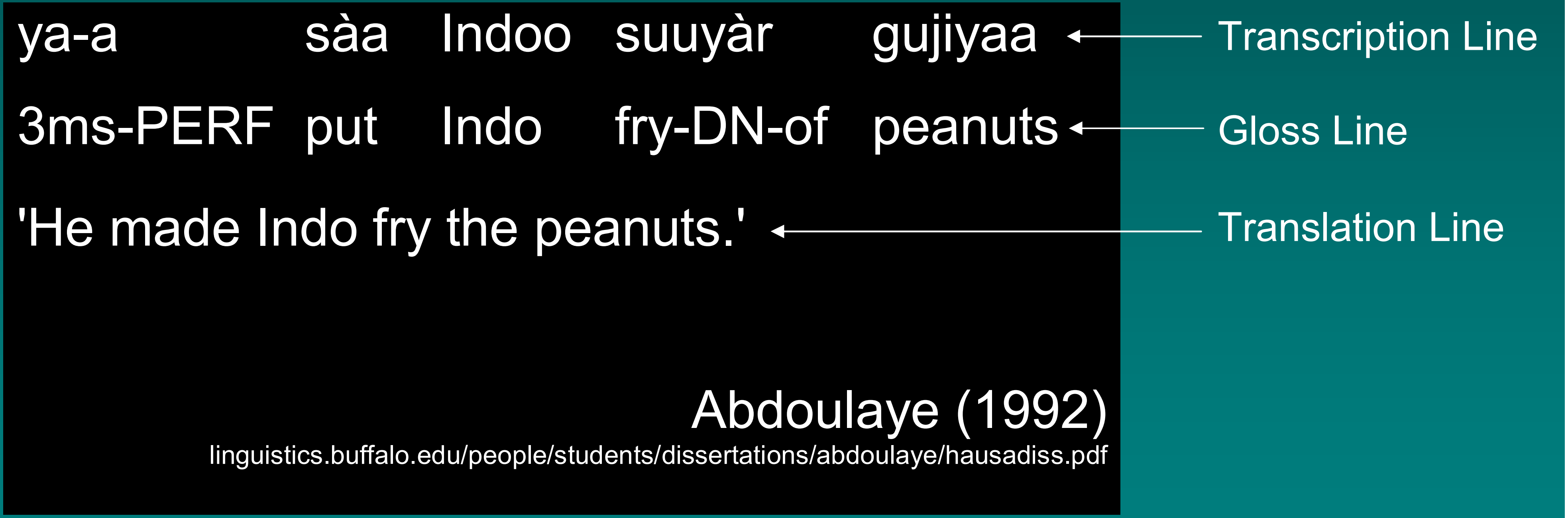
Outline

- Find, Harvest, and Database IGT
- Language ID
- IGT enrichment - Projections, and their Utility
  - Potential for Query
- Evaluation of the Methodology
  - Against independently developed resources
- Conclusion and future work

Finding, Harvesting and Databasing IGT

Interlinear Glossed Text

- Interlinear Glossed Text (IGT) - enriched language data used for illustrative purposes as part of a larger analysis

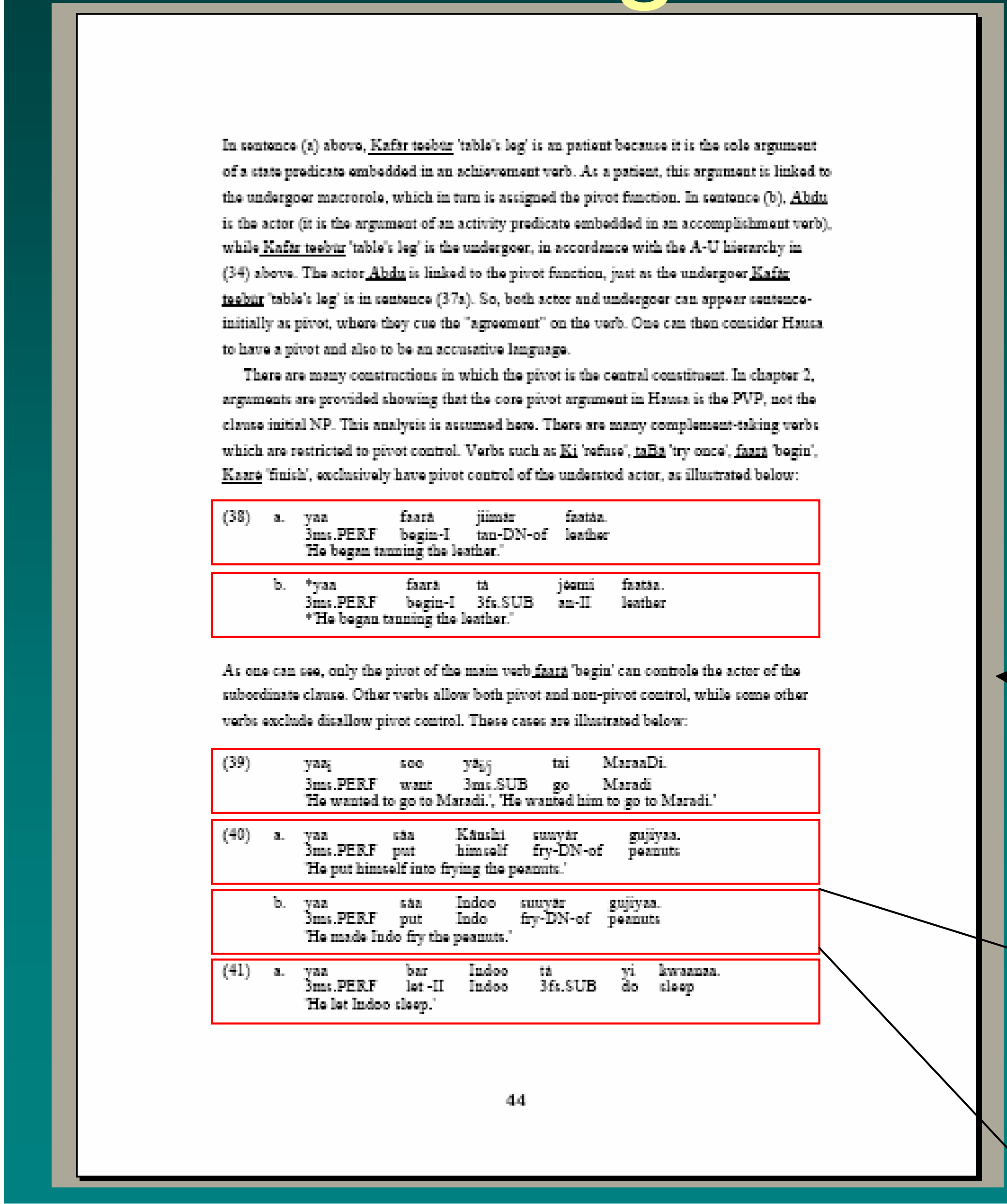


ya-a	sàa	Indoo	suuyàr	gujiyaa
3ms-PERF	put	Indo	fry-DN-of	peanuts

'He made Indo fry the peanuts.'

Abdoulaye (1992)
linguistics.buffalo.edu/people/students/dissertations/abdoulaye/hausadiss.pdf

Locating and Extracting IGT



In sentence (a) above, Kafàr teebùr 'table's leg' is an patient because it is the sole argument of a state predicate embedded in an achievement verb. As a patient, this argument is linked to the undergoer macrorole, which in turn is assigned the pivot function. In sentence (b), Abdu is the actor (it is the argument of an activity predicate embedded in an accomplishment verb), while Kafàr teebùr 'table's leg' is the undergoer, in accordance with the A-U hierarchy in (34) above. The actor Abdu is linked to the pivot function, just as the undergoer Kafàr teebùr 'table's leg' is in sentence (37a). So, both actor and undergoer can appear sentence-initially as pivot, where they cue the "agreement" on the verb. One can then consider Hausa to have a pivot and also to be an accusative language.

There are many constructions in which the pivot is the central constituent. In chapter 2, arguments are provided showing that the core pivot argument in Hausa is the PVP, not the clause initial NP. This analysis is assumed here. There are many complement-taking verbs which are restricted to pivot control. Verbs such as Ki 'refuse', taBa 'try once', faara 'begin', Kaare 'finish', exclusively have pivot control of the understod actor, as illustrated below:

```
(38)   a.   yaa          faarà      jiimàr      faataa.
            3ms.PERF     begin-I    tan-DN-of   leather
            'He began tanning the leather.'
```

```
       b.   *yaa         faara      ta          jeemi     faataa.
            3ms.PERF     begin-I    3fs.SUB     an-II     leather
            *'He began tanning the leather.'
```

As one can see, only the pivot of the main verb faara 'begin' can controle the actor of the subordinate clause. Other verbs allow both pivot and non-pivot control, while some other verbs exclude disallow pivot control. These cases are illustrated below:

```
(39)        yaaᵢ         soo        yaᵢ/ⱼ       tai     MaraaDi.
            3ms.PERF     want       3ms.SUB     go      Maradi
            'He wanted to go to Maradi.', 'He wanted him to go to Maradi.'
```

```
(40)   a.   yaa          sàa      Kânshi     suuyàr      gujiyaa.
            3ms.PERF     put      himself    fry-DN-of   peanuts
            'He put himself into frying the peanuts.'
```

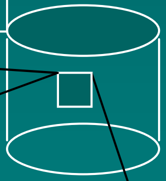
```
       b.   yaa          sàa      Indoo      suuyàr      gujiyaa.
            3ms.PERF     put      Indo       fry-DN-of   peanuts
            'He made Indo fry the peanuts.'
```

```
(41)   a.   yaa          bar      Indoo      ta          yi    kwaanaa.
            3ms.PERF     let -II  Indoo      3fs.SUB     do    sleep
            'He let Indoo sleep.'
```

44

- Find Documents (Crawl)
- Harvest Instances
- Database Instances

Language: Hausa
URL: http://www.ling...



```
ya-a        sàa   Indoo   suuyàr      gujiyaa
3ms-PERF    put   Indo    fry-DN-of   peanuts
'He made Indo fry the peanuts.'
```

Crawling the Web

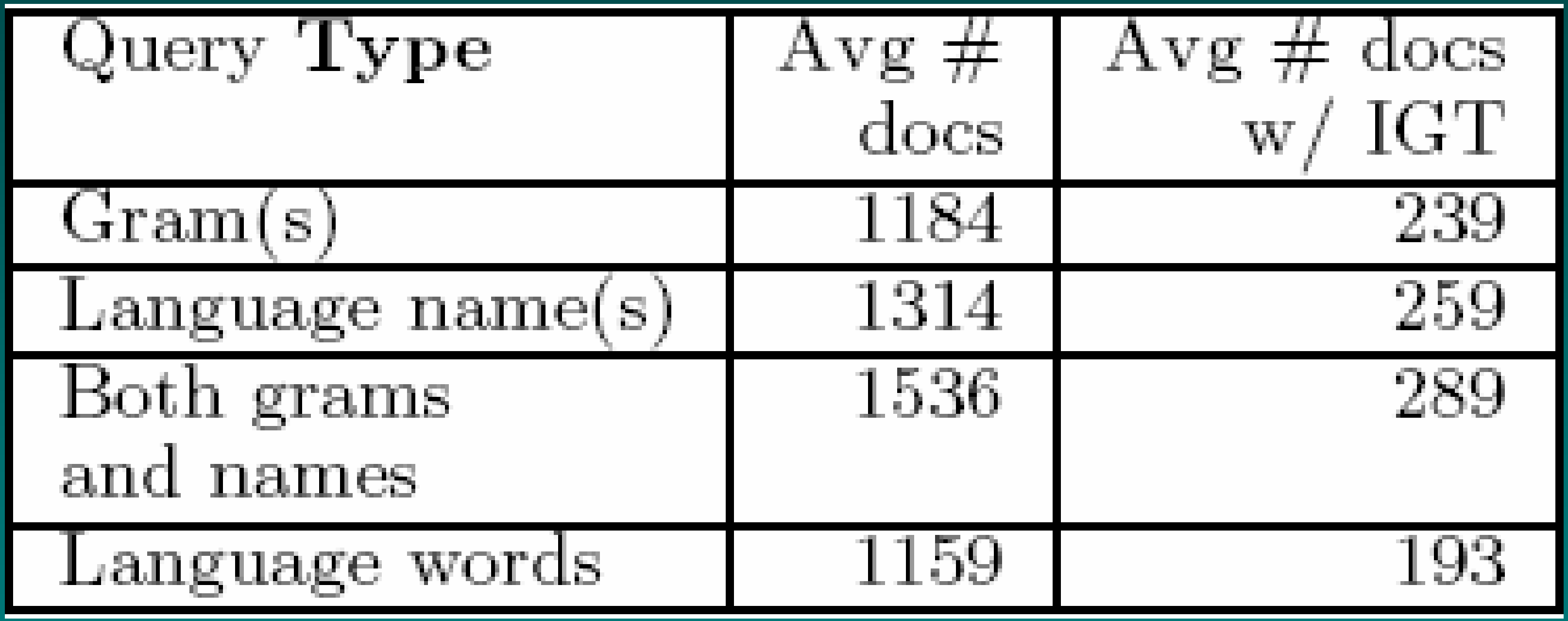
- Intuition: IGT is normally contained in linguistic documents

- Find IGT by throwing queries against existing search engines

- Query terms
  - Grams: -NOM (nominative) , -ACC (accusative)
  - Language names and language codes: Icelandic, Malagasy
    - Drawn from the Ethnologue database (Gordon, 2005)
  - Linguists' names and the languages that they work on:
    - Drawn from the Linguist List's linguist database (linguistlist.org)

- Try different combinations of terms from these categories:
  - Ex: NOM+ACC+Icelandic

Results based on the top 100 queries for each type



Query Type	Avg # docs	Avg # docs w/ IGT
Gram(s)	1184	239
Language name(s)	1314	259
Both grams and names	1536	289
Language words	1159	193

→ “Both grams and names” work best.

IGT detection

- Difficulty in IGT detection
  - Not all IGT are structured the same:
    - Some miss levels of annotation
    - Others add them
    - Some mix annotation within “lines”
    - Long IGT examples are often wrapped multiple times.
  - IGT often embedded in PDFs
    - Pdf-to-text conversion often introduces noise (data loss, corruptions)
    - Encoding not necessarily preserved in extraction - leads to additional data loss

An example

[DP [D0 Ku] [AGRP [Adj ketaran] AGR0 [NP namwu]]]

a.

the big tree

(Kim, 1997)

- •Collapses data & gloss
- •Atypical, “extra” annotations and structure
- •Pdf-to-txt conversion noise

Applying Machine Learning methods to IGT detection

- Treat it as a sequence labeling problem.

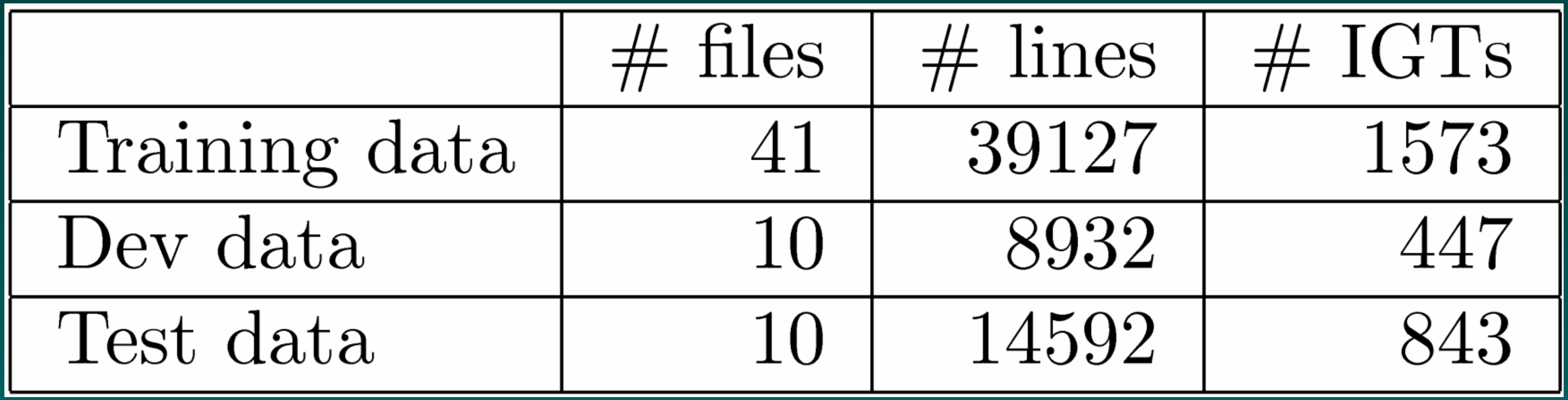
- Label each line in a document with one of the five tags: (an extension of the BIO scheme)
  - BL: a blank line
  - B: the 1st line in an IGT
  - I: inside an IGT that is not a BL
  - E: the last line in an IGT
  - O: outside IGT that is not a BL

- Convert a tag sequence into IGT sequences by simple heuristics:
  - Ex: Any “B [I | BL]* E” sequence is treated as an IGT instance.

Features

- F1: the words that appear on the current line.
- F2: 16 features that look at various cues:
  - Ex: whether the line contains an example number
- F3: the tags of previous two words
- F4: the same as F2 features, but checked against the neighboring lines
  - Ex: whether the next line contains an example number.

Data sets

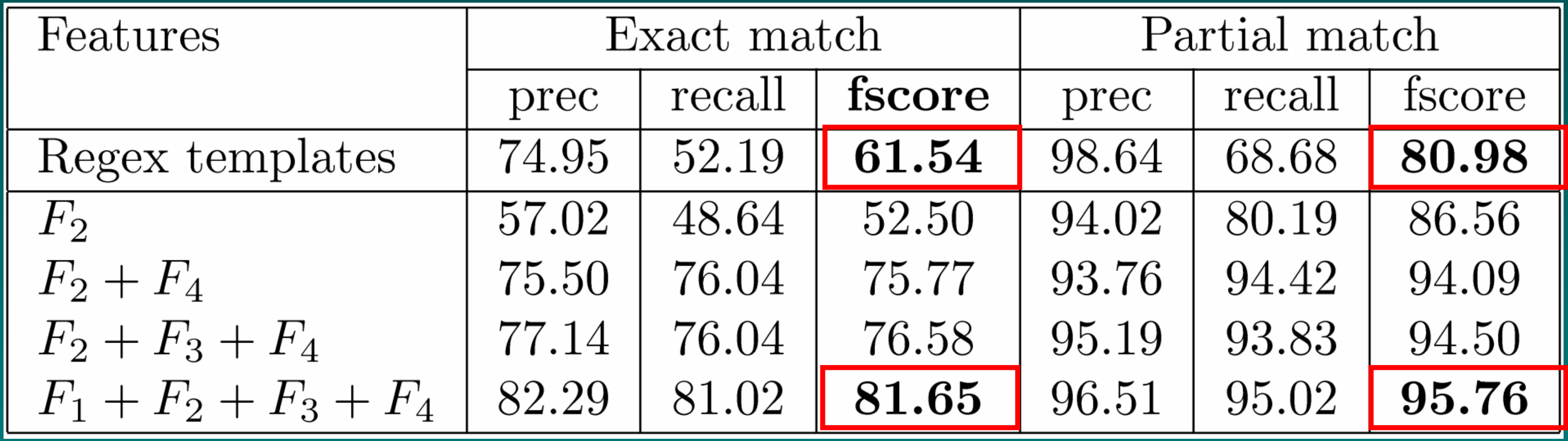


	# files	# lines	# IGTs
Training data	41	39127	1573
Dev data	10	8932	447
Test data	10	14592	843

Evaluation measures:

- Exact match
- Partial match

Performance on the test data



Features	Exact match			Partial match		
	prec	recall	**fscore**	prec	recall	fscore
Regex templates	74.95	52.19	**61.54**	98.64	68.68	**80.98**
F_2	57.02	48.64	52.50	94.02	80.19	86.56
$F_2 + F_4$	75.50	76.04	75.77	93.76	94.42	94.09
$F_2 + F_3 + F_4$	77.14	76.04	76.58	95.19	93.83	94.50
$F_1 + F_2 + F_3 + F_4$	82.29	81.02	**81.65**	96.51	95.02	**95.76**

See Xia & Lewis, IJCNLP 2008

Databasing IGT

- Currently, we parse IGT into a consistent form, stored line-by-line
- We also parse and align glosses with language data
- We POS-tag and parse the English, and provide some search facility over enrichments
- Intuition: IGT are bitexts+
  - We can enrich them further
- And we do language ID and store ISO lang code

Language ID

Language ID

- Language ID essential
  - For query, linguists will insist on it
  - For tool building, incorrect ID can introduce noise
- But…
  - Language ID in IGT is not easy

Previous work on language ID

(not exhaustive)

- (Cavnar and Trenkle, 1994)
- (Damashek, 1995)
- (Elworthy, 1998)
- (Aslam and Frost, 2003)
- (McNamee and Mayfield, 2004)
- (Kruengkrai et al., 2005)
- ….

A good summary in (Hughes et. al., 2006)

They all require a reasonable amount of training data for each language.

Differences from a typical language ID task

- Large number of languages: 600+
- Unseen languages: 10% of IGTs in test data belong to unseen languages
- Very limited amount of training data: no more than 10 words per language for 45.3% of languages
- …

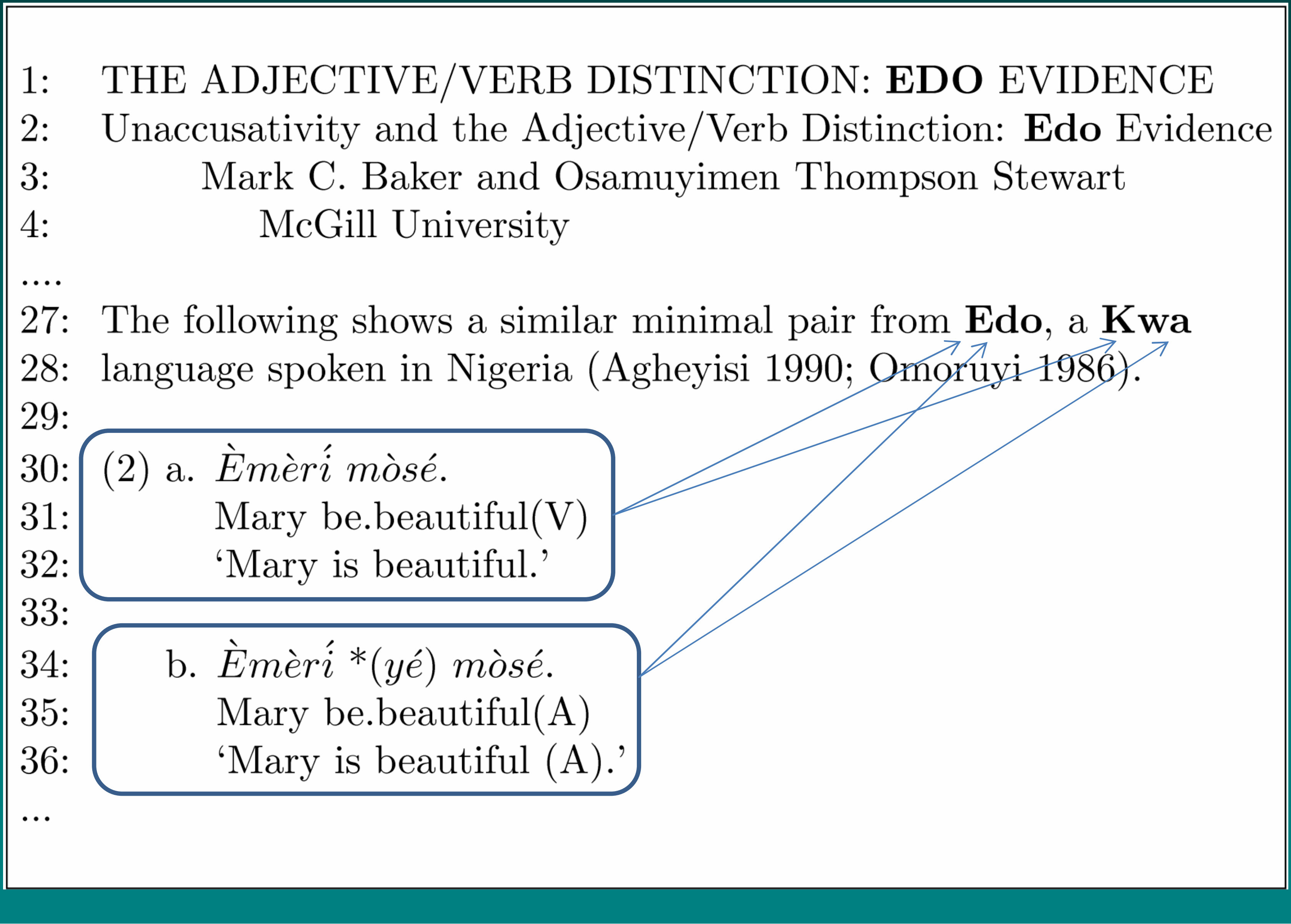
➔Cavnar and Trenkle’s algorithm: 99.8% (8 langs)

➔For us (600+ languages) => C&T returns 51.4%

Use of language code

- A language can have multiple names:
  - Ex: “aaa” => Alumu, Tesu, Arum, Alumu-Tesu, Alumu, Arum-Cesu, Arum-Chessu, and Arum-Tesu
- A language name can refer to multiple languages:
  - Ex: Edo => “bin” or “lew”
- We use language codes, because each language code maps to exactly one language
- Our system outputs both language codes and language names

Language ID



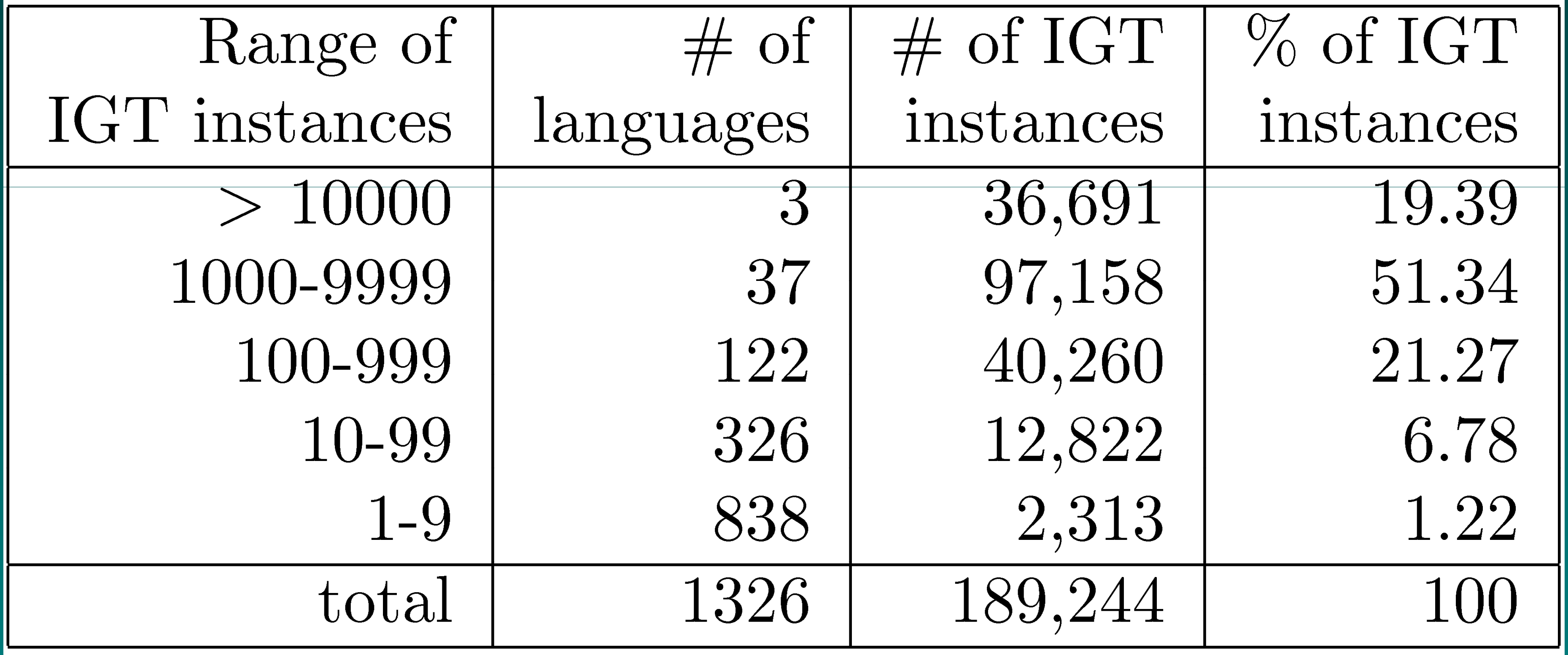
1: THE ADJECTIVE/VERB DISTINCTION: **EDO** EVIDENCE
2: Unaccusativity and the Adjective/Verb Distinction: **Edo** Evidence
3: Mark C. Baker and Osamuyimen Thompson Stewart
4: McGill University
....
27: The following shows a similar minimal pair from **Edo**, a **Kwa**
28: language spoken in Nigeria (Agheyisi 1990; Omoruyi 1986).
29:
30: (2) a. *Èmèrí mòsé.*
31: Mary be.beautiful(V)
32: 'Mary is beautiful.'
33:
34: b. *Èmèrí *(yé) mòsé.*
35: Mary be.beautiful(A)
36: 'Mary is beautiful (A).'
...

Language ID (cont)

- Standard language ID algorithms do not work
  - Large number of languages
  - Little training data
  - …

- Our work:
  - Treating language ID as a co-reference task
    - Mary called Chris. She was running late.
  - Applying NLP techniques (e.g., MaxEnt, Markov logic, etc.)
  - Results (in accuracy): 85.10%

ODIN database



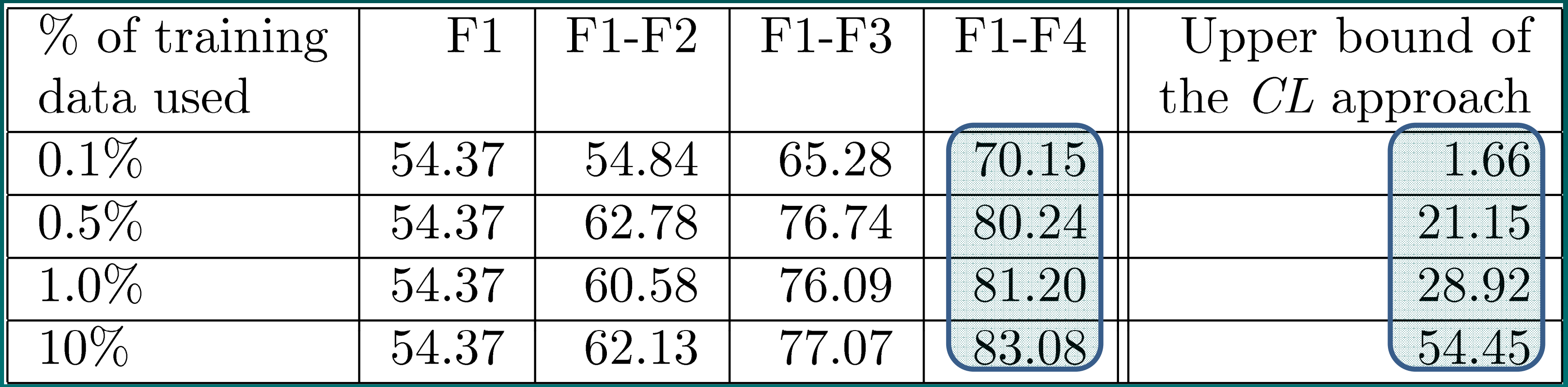
Range of IGT instances	# of languages	# of IGT instances	% of IGT instances
> 10000	3	36,691	19.39
1000-9999	37	97,158	51.34
100-999	122	40,260	21.27
10-99	326	12,822	6.78
1-9	838	2,313	1.22
total	1326	189,244	100

Feature templates

- (F1) The nearest language that precedes the IGT
- (F2) The languages appearing in the neighborhood of the IGT
- (F3) Comparing ngrams in the current IGT and ngrams for a language

  => This is info used in a traditional language ID algorithm
- (F4) Comparing ngrams in the current IGT and ngrams in other IGTs in the same document

With less training data



% of training data used	F1	F1-F2	F1-F3	F1-F4	Upper bound of the *CL* approach
0.1%	54.37	54.84	65.28	70.15	1.66
0.5%	54.37	62.78	76.74	80.24	21.15
1.0%	54.37	60.58	76.09	81.20	28.92
10%	54.37	62.13	77.07	83.08	54.45

See Xia, Lewis, & Poon, EACL 2009

Where we are

- Online
  - ODIN has 41,545 instance collected from 2,946 documents
  - All collected from the original regex approach
  - 45% hand reviewed
- Soon to come online
  - 189,000+ instances identified using the new ML techniques *from the same documents*
  - Most have been hand reviewed
- In the near future
  - 100,000+ documents have been identified that *might* contain IGT (crawling continues unabated)
  - All of these documents will be run through the new tools and added
  - Anticipate 500K-1M+ new instances of IGT
- Unifying markup
  - Limited, mostly manual, work thus far
  - Targeted for future ML work
- Correcting instances (fixing noise)
  - Another application of ML technology (heuristics only get us so far)

Enriching IGT

Main Ideas

- Project annotations and structures onto target language data
  - Structures include
    - Annotations
    - Dependency structures
    - Phrase structures
- Process could be used to normalize annotations used in the database (to facilitate search)

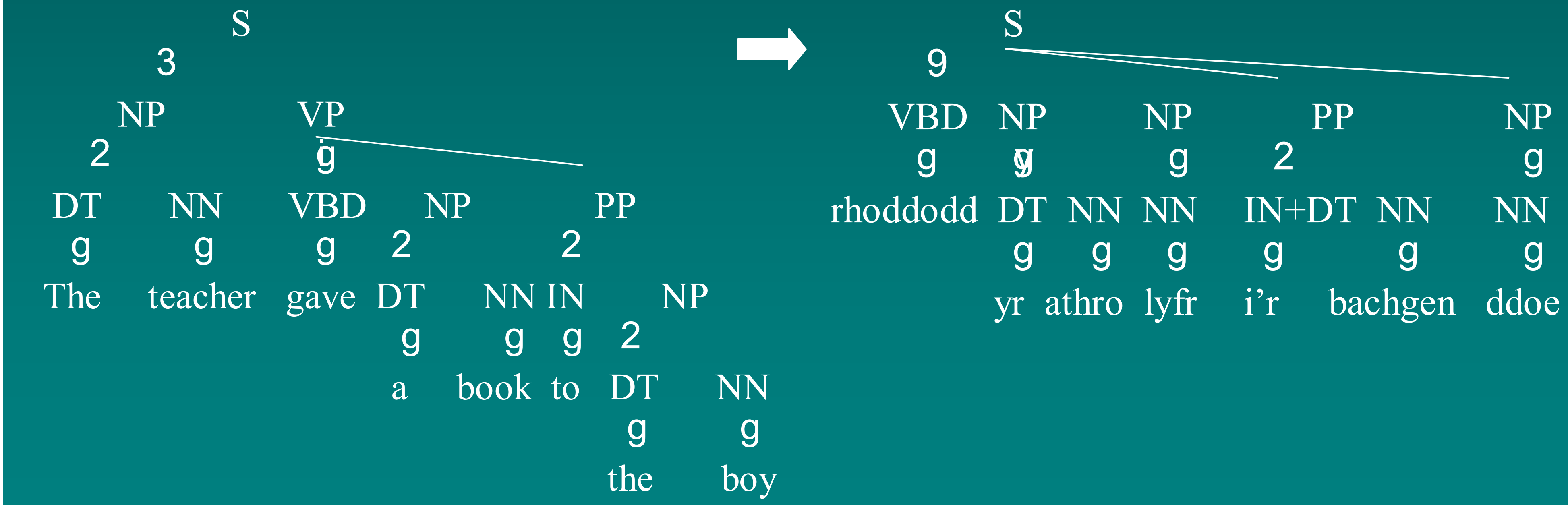
Projection

Enriched English data

The teacher gave a book to the boy
DT NN VBD DT NN IN DT NN

Welsh language data

Rhoddodd yr athro lyfr i’r bachgen
VBD DT NN NN IN-DT NN



Structural projection work

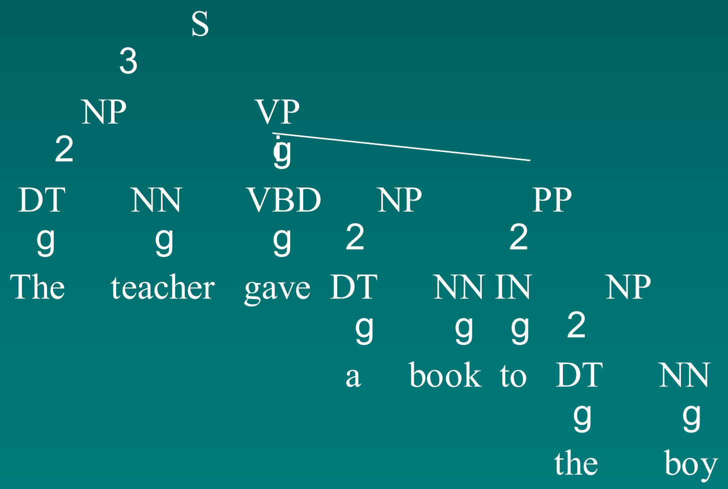
- Previous work
  - (Yarowsky & Ngai, 2001): POS tags and NP boundaries
  - (Xi & Hwa, 2005): POS tags
  - (Hwa et al., 2002): dependency structures
  - (Quirk et al., 2005): dependency structures

- Current projection work:
  - Projecting both dependency structures (Lewis et al 2006)
  - …and phrase structures (Xia and Lewis 2007)
  - Does not require a large amount of parallel data or hand-aligned data for accurate projections
  - Can be applied to hundreds of languages, drawing from ODIN (Lewis 2006)

Some Notes

Notations and Terminology

- Part of Speech labels use Penn Treebank (PTB) tags
  - E.g., DT=determiner, NN=noun, VB=verb, etc.
- Trees use PTB phrasal labels (~GB) & non-binary branching



- “Projections” ≠ “syntactic projections” (as in the EPP, Chomsky 1981)

The Methodology

- For the IGT for any language:

1. Parse the English translation to produce a syntactic tree
2. Align the target language data and the translation, notably through the gloss line
3. Project annotations and the syntactic tree onto the target language data
4. Reorder tree according to linear order of the constituents in the target sentence

Sample IGT Instance

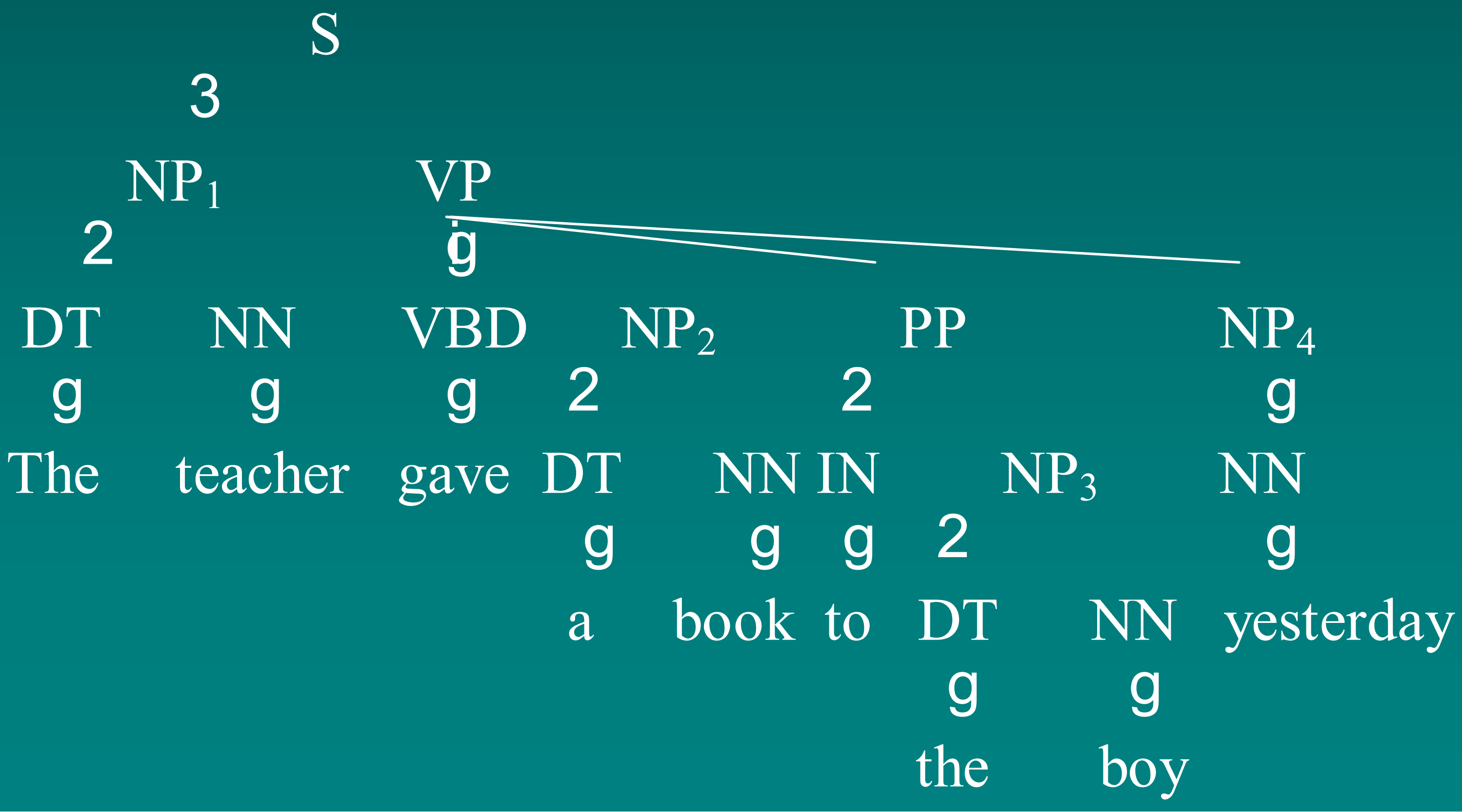
Rhoddodd yr athro lyfr i’r bachgen ddoe
Gave-**3sg** the teacher book to-the boy yesterday
“The teacher gave a book to the boy yesterday”

(Bailyn, 2001)

Step 1 - Parse

- Parse the English translation (e.g. using Charniak’s parser, Charniak 97, or Collin’s parser, Collins 98):

“The teacher gave a book to the boy yesterday.”



Step 2: Word alignment

- Align the translation with the target:

Rhoddodd	yr	athro	lyfr	i’r	bachgen	ddoe
gave-3sg	the	teacher	book	to-the	boy	yesterday

“The teacher gave a book to the boy yesterday”

Step 2: Word alignment

- Align the translation with the target:

Rhoddodd	yr	athro	lyfr	i'r	bachgen	ddoe
gave-3sg	the	teacher	book	to-the	boy	yesterday

"The teacher gave a book to the boy yesterday"

Step 2: Word alignment

- Align the translation with the target:

Rhoddodd	yr	athro	lyfr	i’r	bachgen	ddoe
gave-3sg	the	teacher	book	to-the	boy	yesterday

“The teacher gave a book to the boy yesterday”

Step 2: Word alignment

- Align the translation with the target:

Rhoddodd	yr	athro	lyfr	i’r	bachgen	ddoe
gave-3sg	the	teacher	book	to-the	boy	yesterday

“The teacher gave a book to the boy yesterday”

Step 2: Word alignment

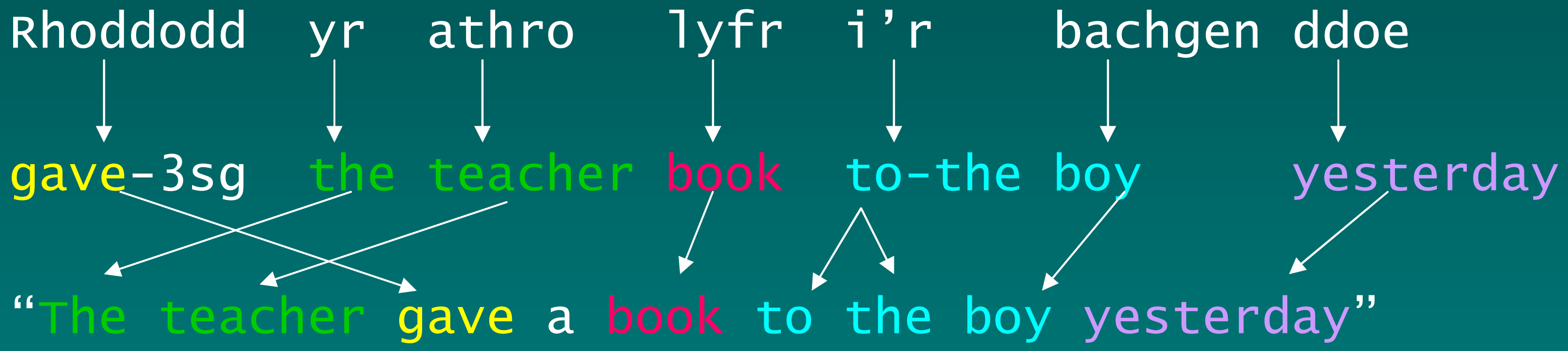
- Align the translation with the target:

Rhoddodd	yr	athro	lyfr	i’r	bachgen	ddoe
gave-3sg	the	teacher	book	to-the	boy	yesterday

“The teacher gave a book to the boy yesterday”

Step 2: Word alignment

- Align the translation with the target:



Rhoddodd yr athro lyfr i'r bachgen ddoe

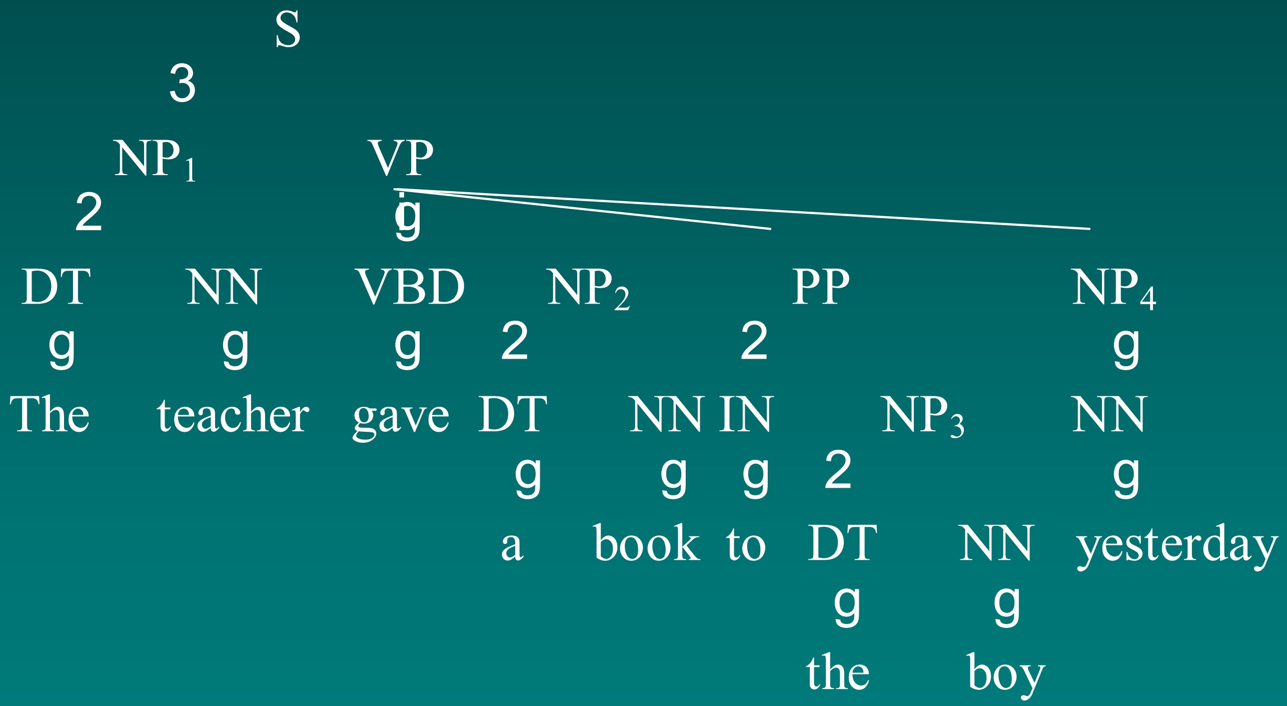
gave-3sg the teacher book to-the boy yesterday

"The teacher gave a book to the boy yesterday"

Step 3 – Project Structure

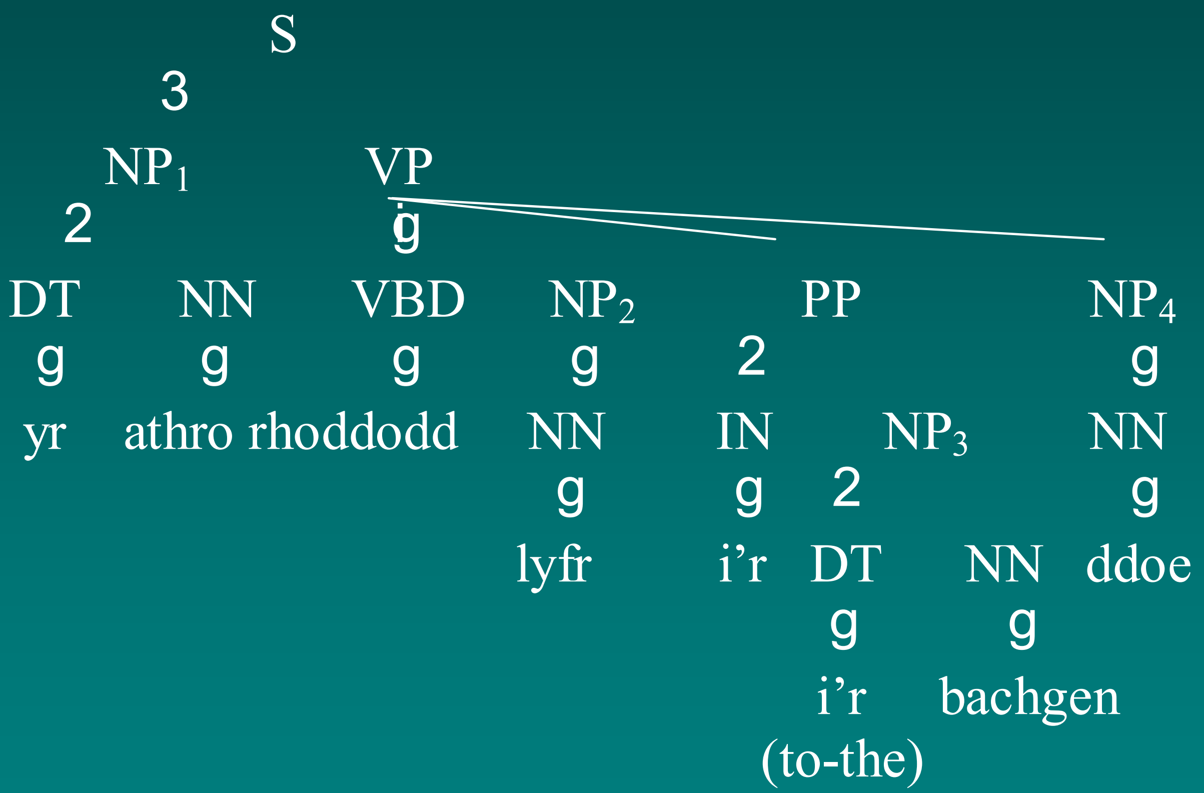
- Copy the English tree and remove all the unaligned English words
- Replace English words with corresponding target words
- Remove duplicates (if any) and attach unaligned target words
- Reorder tree (according to the linear order of the target)

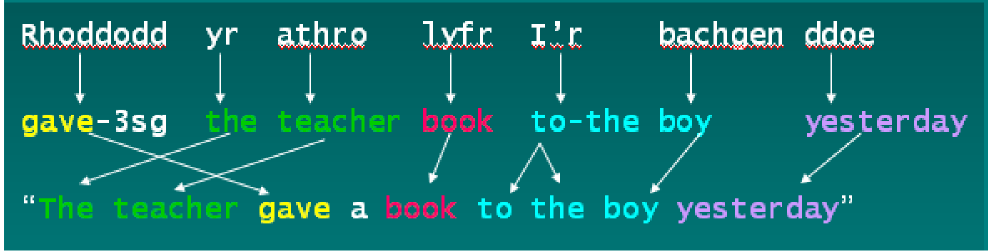
Start with English tree



“The teacher gave a book to the boy yesterday”

Replace English words with target words



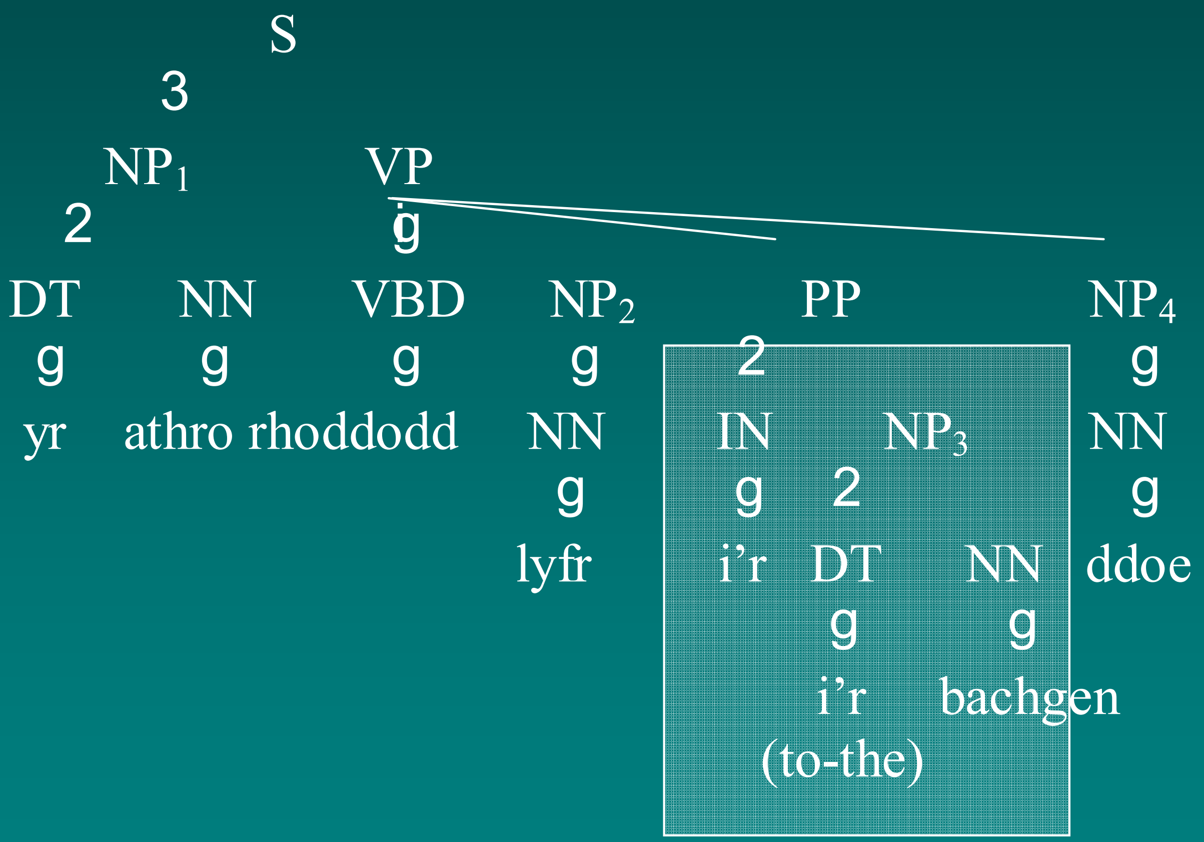


Rhoddodd yr athro lyfr I’r bachgen ddoe

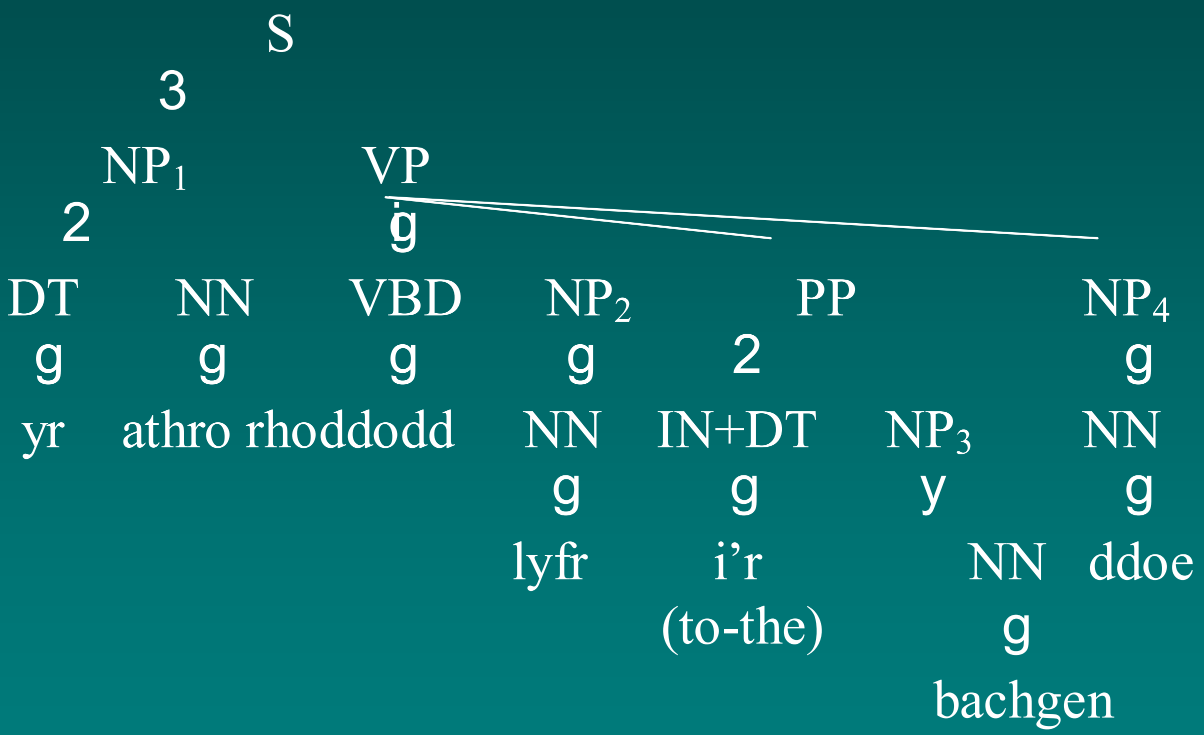
gave-3sg the teacher book to-the boy yesterday

“The teacher gave a book to the boy yesterday”

Remove Duplicates



Remove Duplicates



S

NP_1 VP

DT NN VBD NP_2 PP NP_4

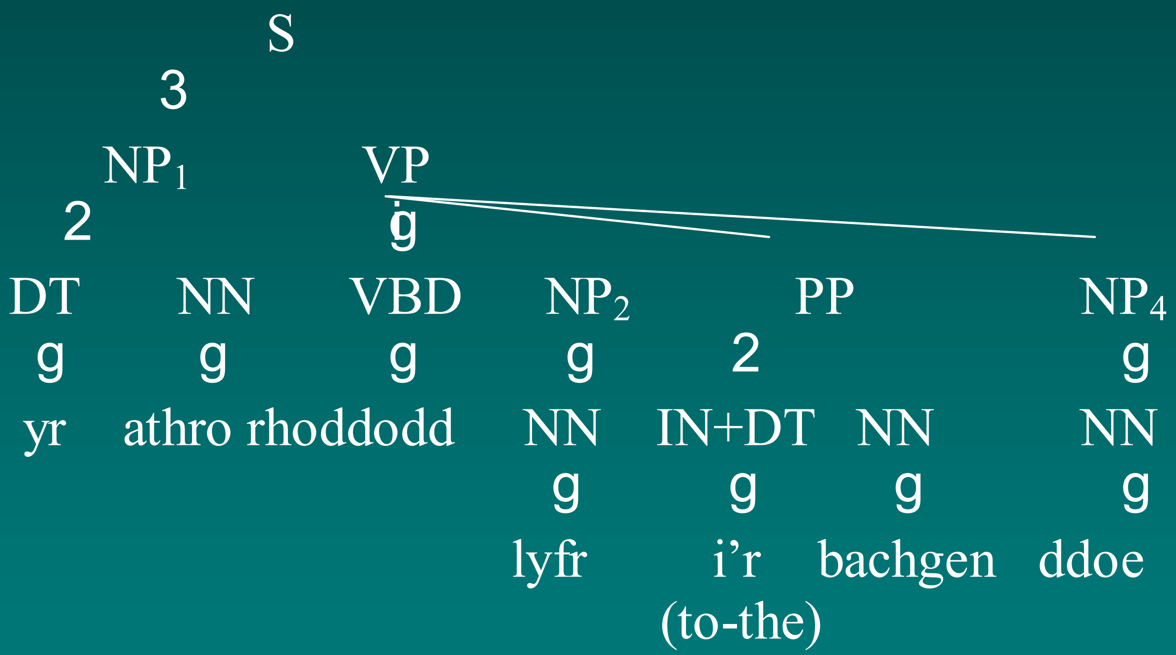
yr athro rhoddodd NN IN+DT NP_3 NN

lyfr i’r NN ddoe

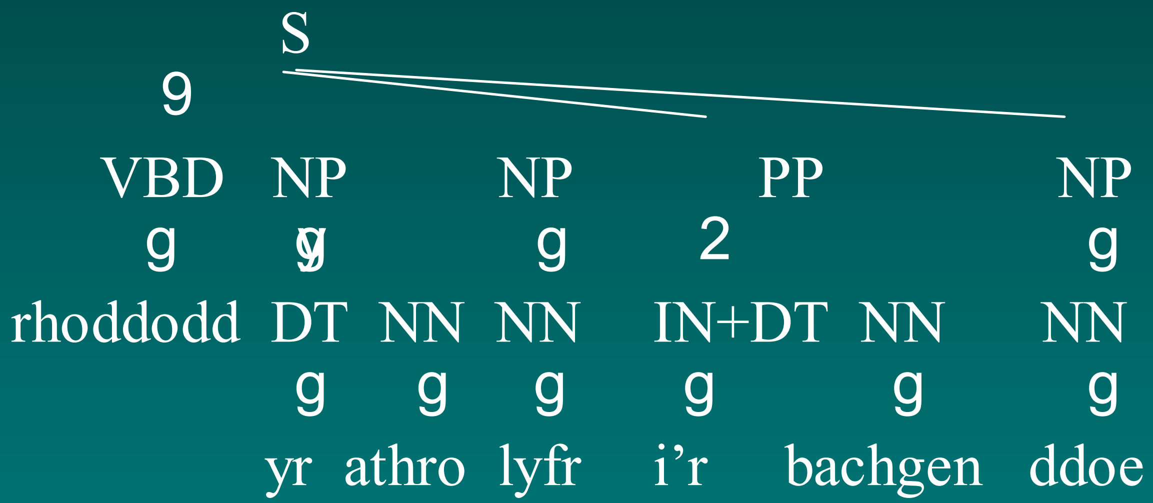
(to-the)

bachgen

Remove Duplicates



Step 4 - Reorder



S

9

VBD NP NP PP NP

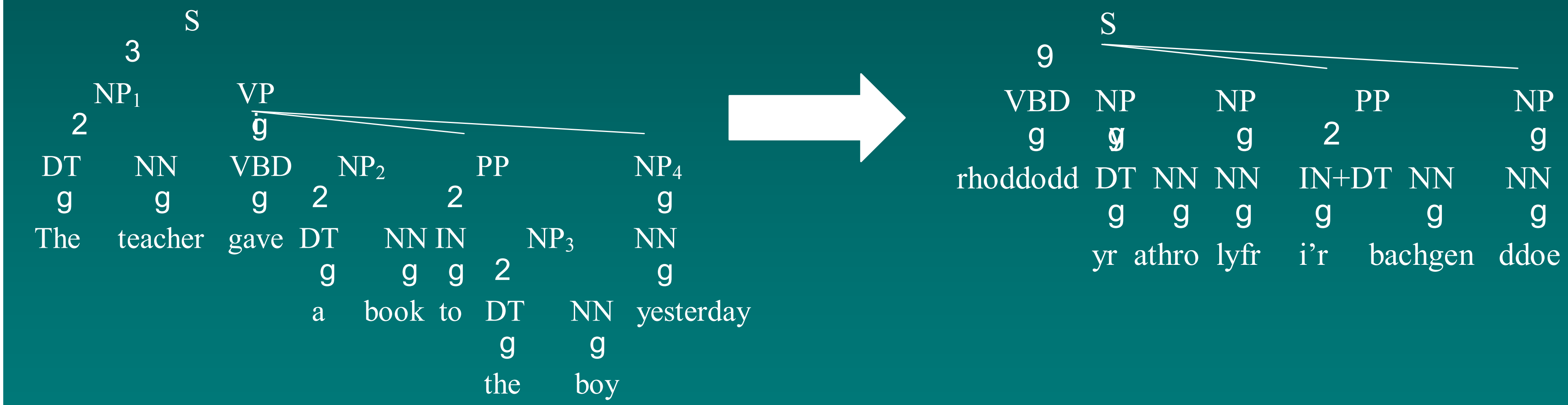
g g g 2 g

rhoddodd DT NN NN IN+DT NN NN

g g g g g g

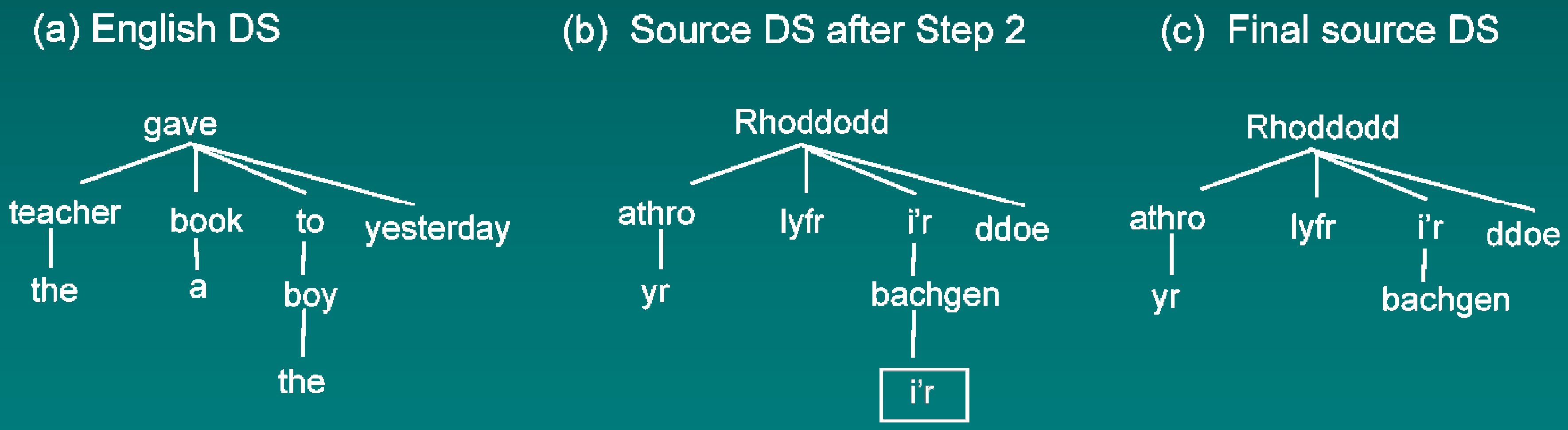
yr athro lyfr i’r bachgen ddoe

Summary of the projection algorithm



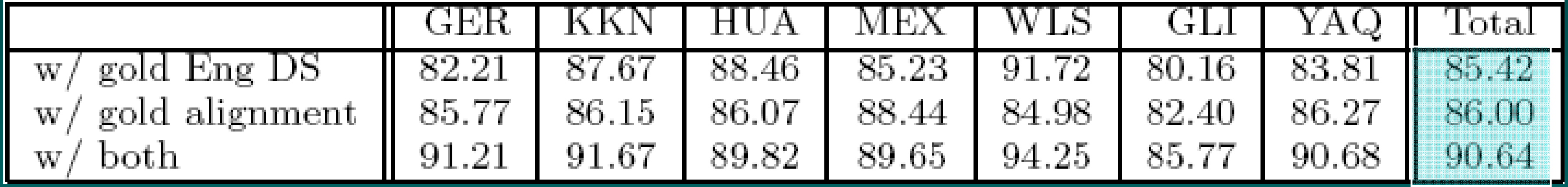
Dependency Structure Projection

- We also can build and project dependency structures:



Projection results (DS only)

- Results from Lewis et al 2006:



	GER	KKN	HUA	MEX	WLS	GLI	YAQ	Total
w/ gold Eng DS	82.21	87.67	88.46	85.23	91.72	80.16	83.81	85.42
w/ gold alignment	85.77	86.15	86.07	88.44	84.98	82.40	86.27	86.00
w/ both	91.21	91.67	89.82	89.65	94.25	85.77	90.68	90.64

(Measured against gold standards created by human annotators.)

Utility of Projections

Construction Query

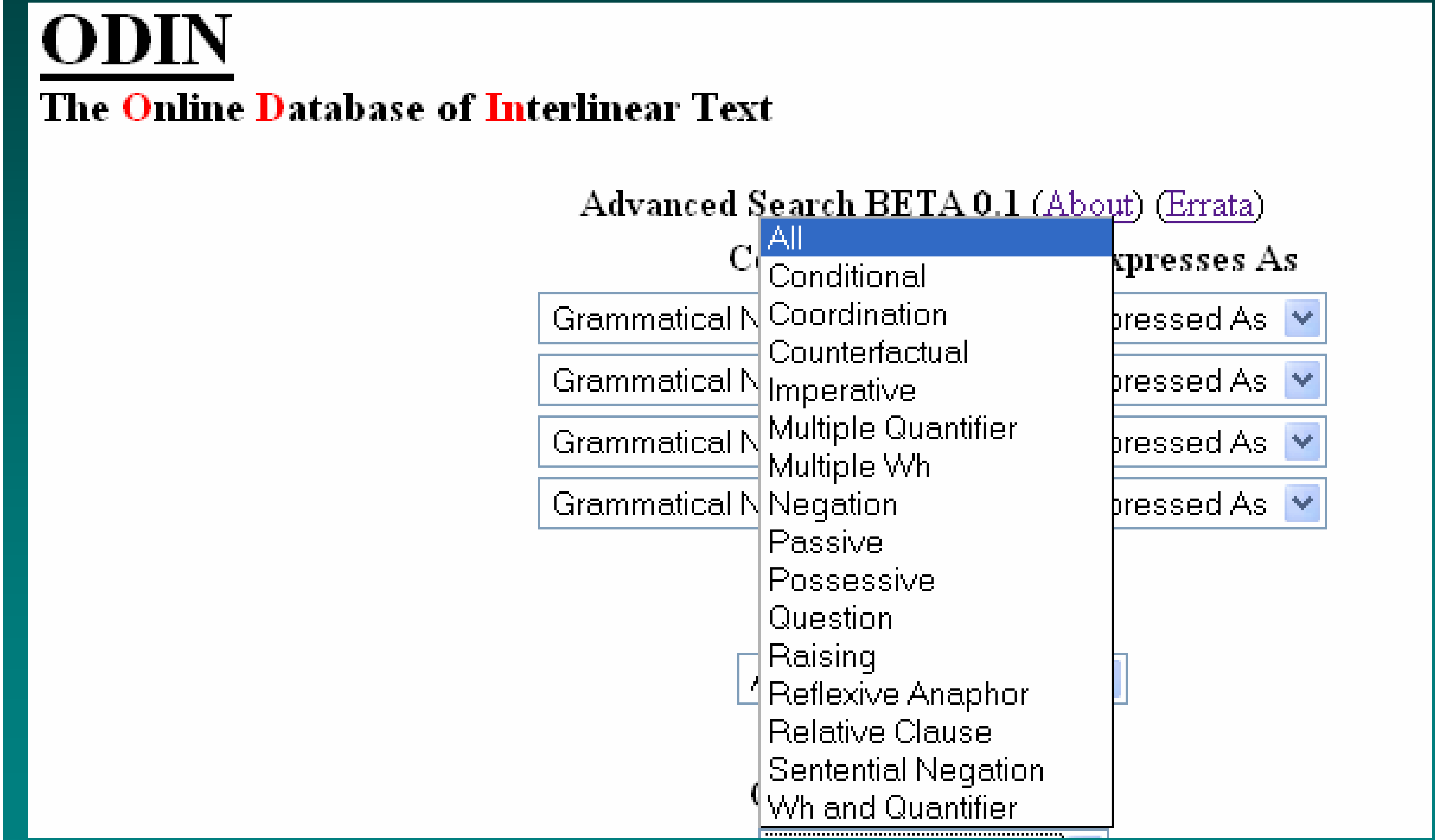
Construction Query

- Question:
  - Can we search *cross-linguistically* for constructions based on syntactic or morphosyntactic cues?
- Assumption:
  - There are universal constructions *and*
  - There are syntactic or morphosyntactic reflexes of these constructions.

Construction Query

- Given annotated and parsed English data, we can
- Search for constructions like:
  - Passives
  - Relative clauses
  - Raising constructions
  - Sluices
  - Focus (English “It’s the xx that”)
- The aligned language data in IGT *might* contain similar constructions

ODIN Construction Query

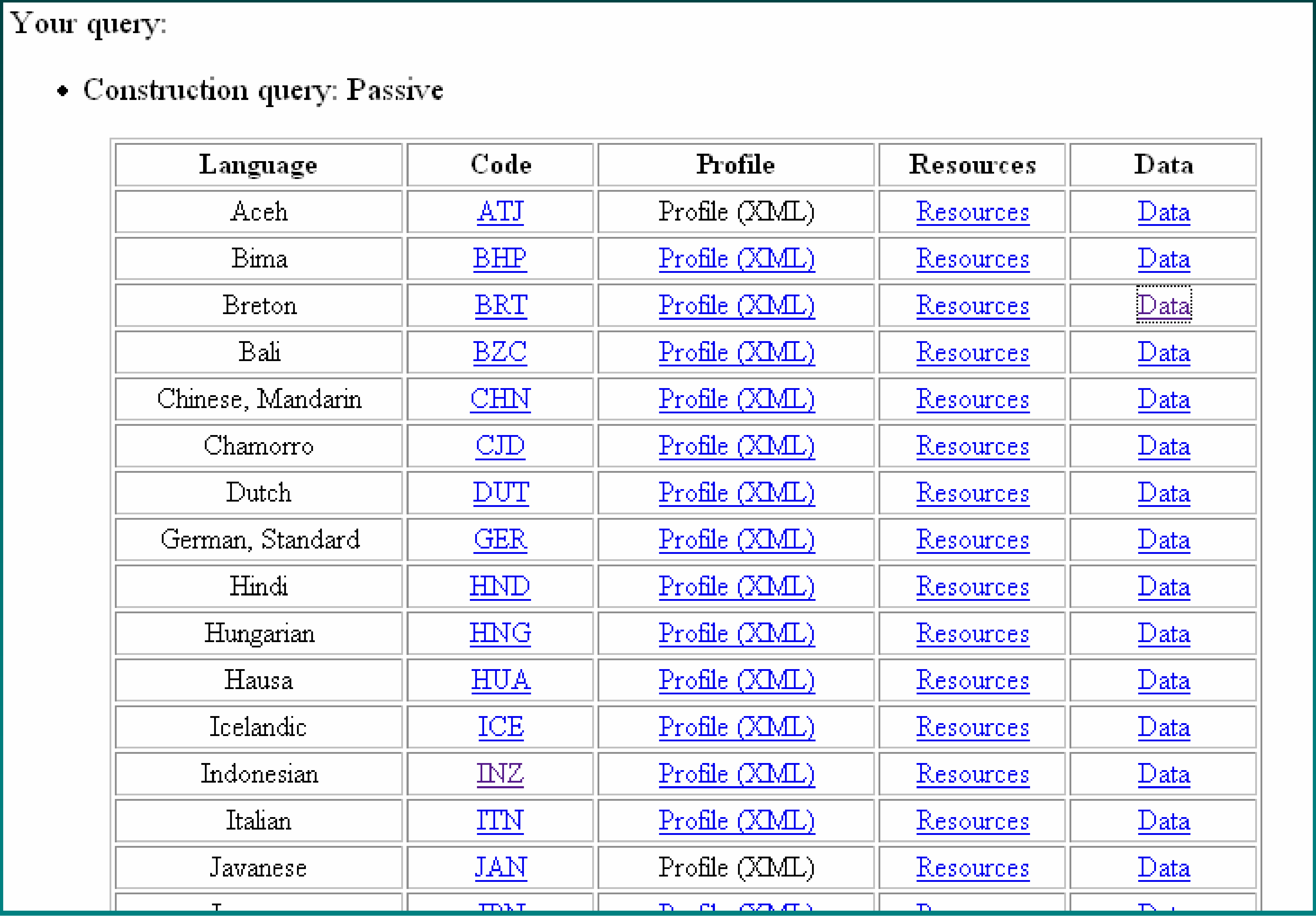


ODIN

The Online Database of Interlinear Text

Advanced Search BETA 0.1 (About) (Errata)

Langs w/ passive examples (maybe)

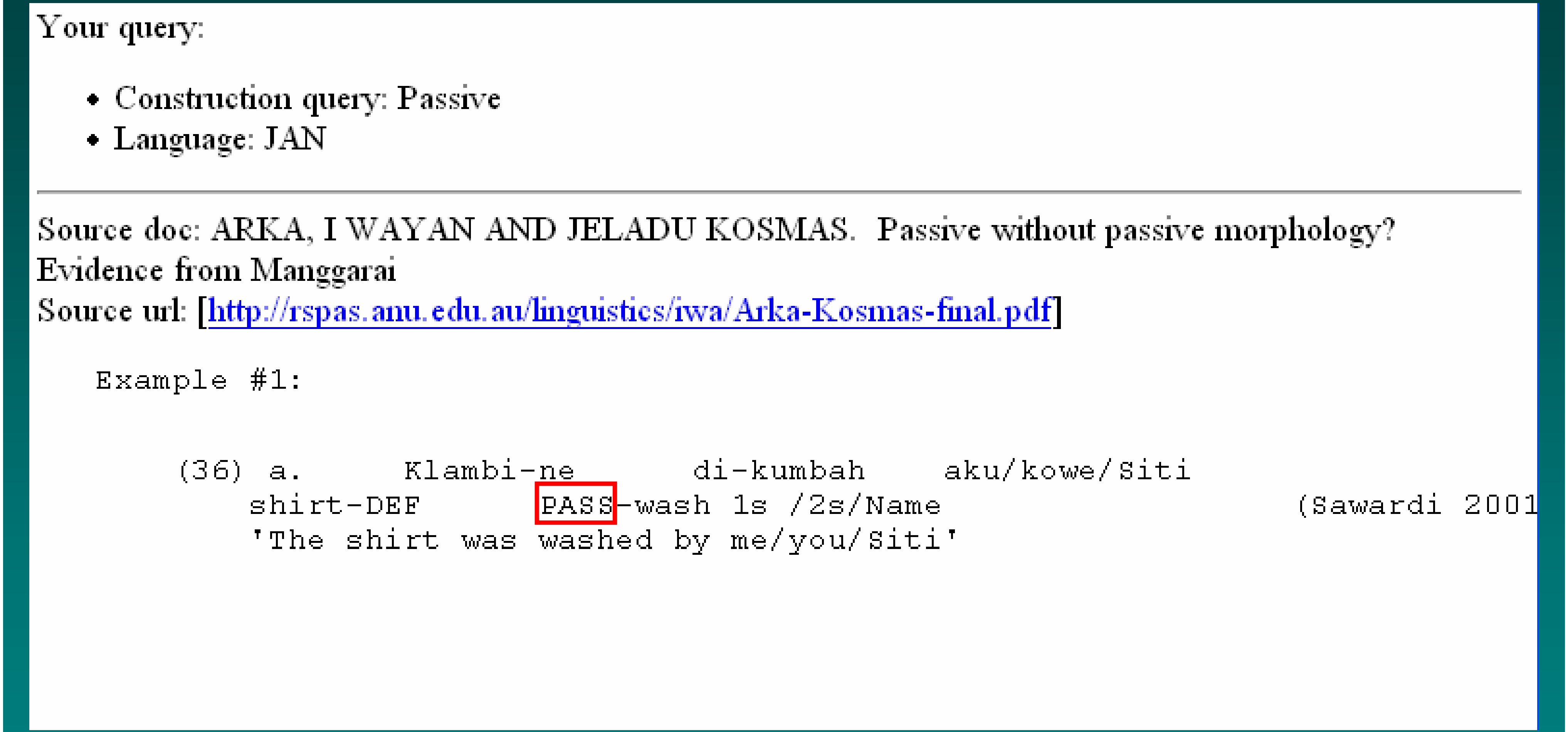


Your query:

- Construction query: Passive

Language	Code	Profile	Resources	Data
Aceh	ATJ	Profile (XML)	Resources	Data
Bima	BHP	Profile (XML)	Resources	Data
Breton	BRT	Profile (XML)	Resources	Data
Bali	BZC	Profile (XML)	Resources	Data
Chinese, Mandarin	CHN	Profile (XML)	Resources	Data
Chamorro	CJD	Profile (XML)	Resources	Data
Dutch	DUT	Profile (XML)	Resources	Data
German, Standard	GER	Profile (XML)	Resources	Data
Hindi	HND	Profile (XML)	Resources	Data
Hungarian	HNG	Profile (XML)	Resources	Data
Hausa	HUA	Profile (XML)	Resources	Data
Icelandic	ICE	Profile (XML)	Resources	Data
Indonesian	INZ	Profile (XML)	Resources	Data
Italian	ITN	Profile (XML)	Resources	Data
Javanese	JAN	Profile (XML)	Resources	Data

Passive examples (maybe)



Your query:

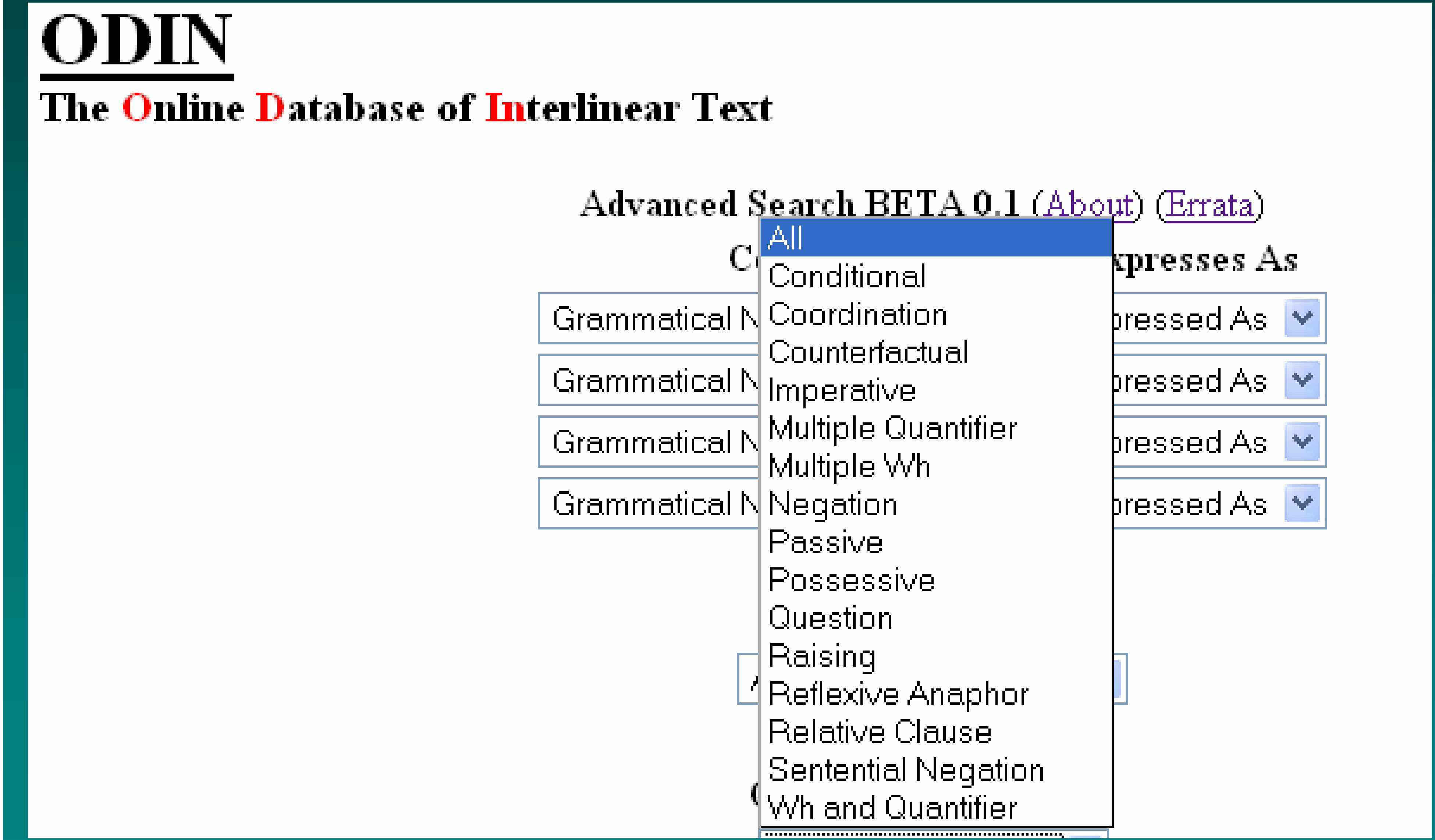
- Construction query: Passive
- Language: JAN

Source doc: ARKA, I WAYAN AND JELADU KOSMAS. Passive without passive morphology? Evidence from Manggarai
Source url: [http://rspas.anu.edu.au/linguistics/iwa/Arka-Kosmas-final.pdf]

```
   Example #1:


       (36) a.     Klambi-ne         di-kumbah     aku/kowe/Siti
            shirt-DEF        PASS-wash 1s /2s/Name                  (Sawardi 2001
            'The shirt was washed by me/you/Siti'
```

ODIN Construction Query

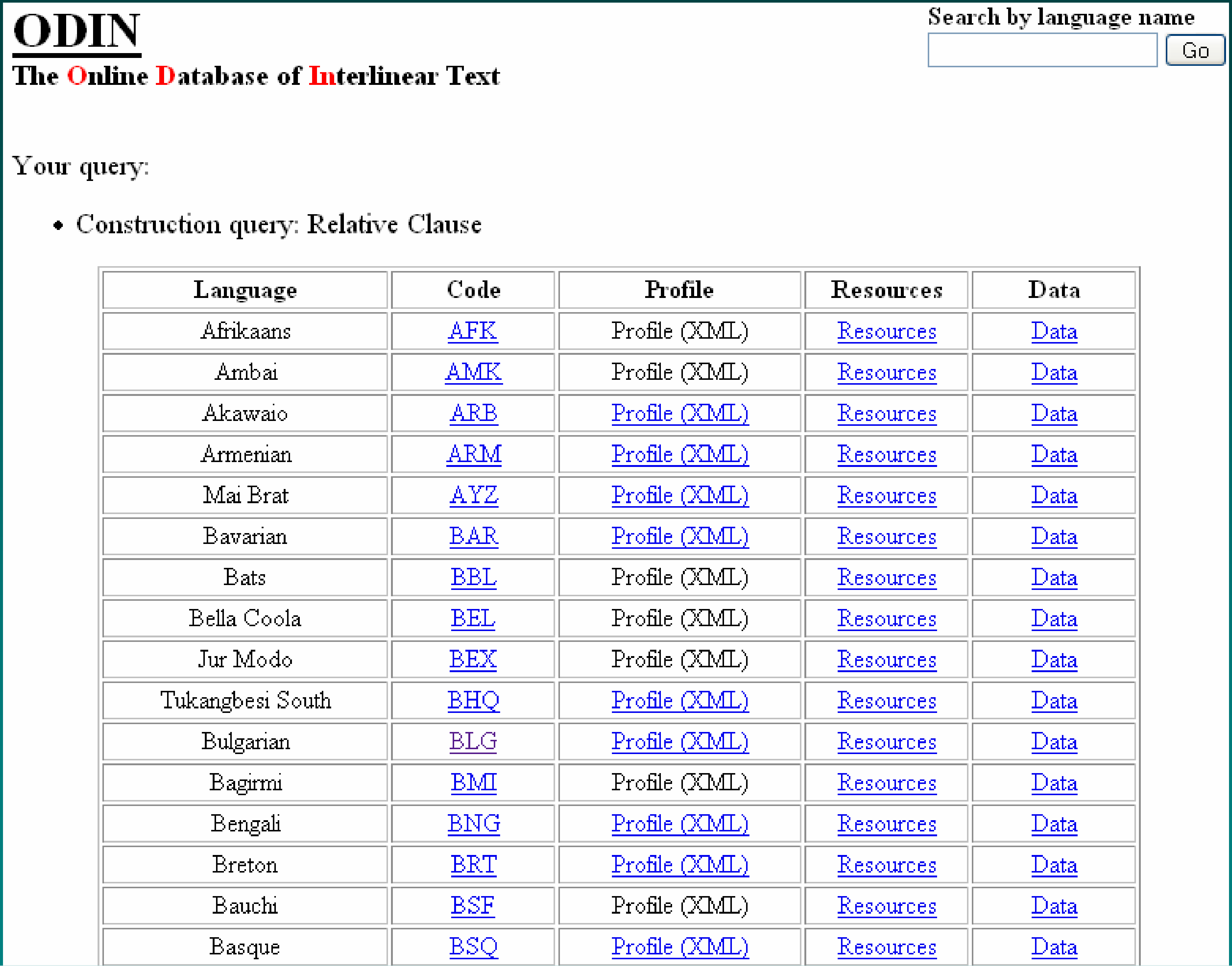


ODIN

The Online Database of Interlinear Text

Advanced Search BETA 0.1 (About) (Errata)

Langs w/ relative clauses (maybe)



ODIN

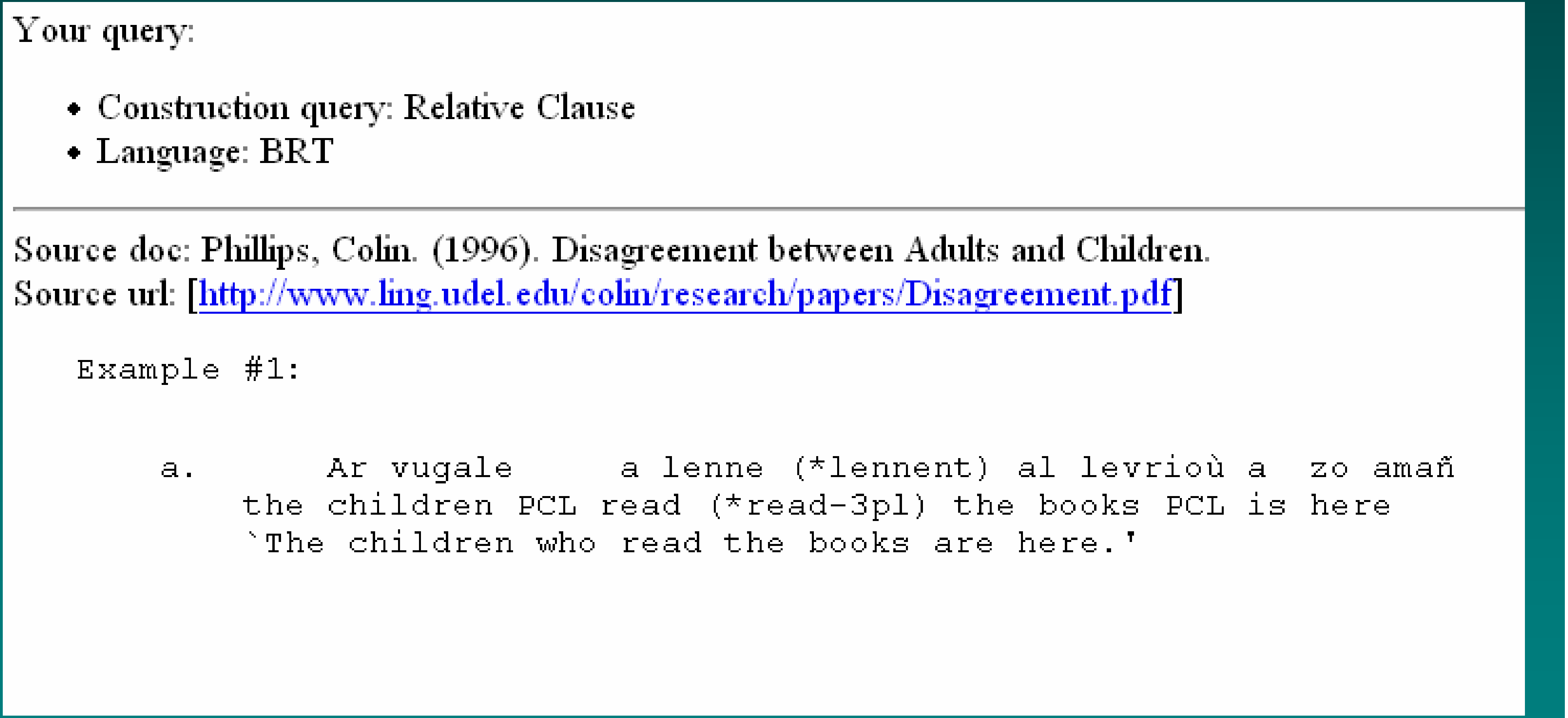
The Online Database of Interlinear Text

Your query:

- Construction query: Relative Clause

Language	Code	Profile	Resources	Data
Afrikaans	AFK	Profile (XML)	Resources	Data
Ambai	AMK	Profile (XML)	Resources	Data
Akawaio	ARB	Profile (XML)	Resources	Data
Armenian	ARM	Profile (XML)	Resources	Data
Mai Brat	AYZ	Profile (XML)	Resources	Data
Bavarian	BAR	Profile (XML)	Resources	Data
Bats	BBL	Profile (XML)	Resources	Data
Bella Coola	BEL	Profile (XML)	Resources	Data
Jur Modo	BEX	Profile (XML)	Resources	Data
Tukangbesi South	BHQ	Profile (XML)	Resources	Data
Bulgarian	BLG	Profile (XML)	Resources	Data
Bagirmi	BMI	Profile (XML)	Resources	Data
Bengali	BNG	Profile (XML)	Resources	Data
Breton	BRT	Profile (XML)	Resources	Data
Bauchi	BSF	Profile (XML)	Resources	Data
Basque	BSQ	Profile (XML)	Resources	Data

Relative Clause?



Your query:

- Construction query: Relative Clause
- Language: BRT

Source doc: Phillips, Colin. (1996). Disagreement between Adults and Children.
Source url: [http://www.ling.udel.edu/colin/research/papers/Disagreement.pdf]

```
Example #1:


      a.     Ar vugale     a lenne (*lennent) al levrioù a  zo amañ
             the children PCL read (*read-3pl) the books PCL is here
             `The children who read the books are here.'
```

Other queries

- Search English structures and annotations, and their alignments within target language data
  - E.g., Search for relative clauses
  - Does the language use relative pronouns, etc.? (cf Comrie 2006)
- Search enriched target language data directly
  - Constituency
  - Values for typological parameters (specifically structural)
  - Constructions

Concerns

- A database of IGT a great resource, but…
- Issues of reliability with its use for structural projections:
  - IGT bias
    - Tend to be short
    - "Skewed" examples (e.g., scrambled, non-canonical forms, etc.)
  - English bias
    - The source language is English!
    - Projected structures can
      - Contain only enough detail as found in annotated English (and glosses)
      - Annotations, POS tags, phrasal types will all be English-centric
  - Treebank bias
  - Noise
    - PDF Extraction
    - "Faux" IGT

Concerns

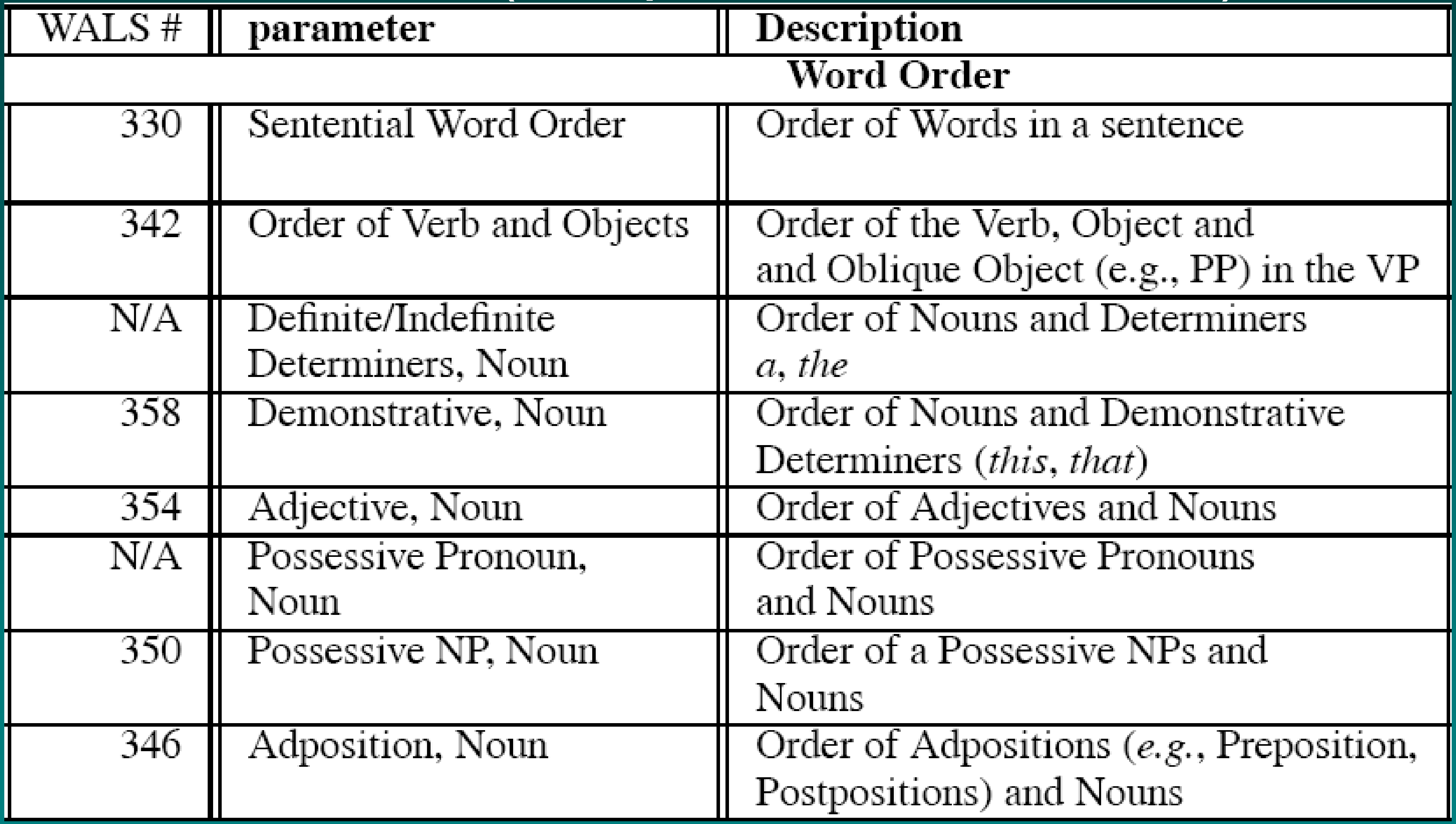
- How much of a problem are the IGT and English biases, really?
- Lewis & Xia (2008): Set of experiments to test:
  1. Utility of projected structures for typological queries (particularly where syntactic structures essential) – English bias
  2. Determine how much data we need to overcome skewed data – IGT bias
- Test empirically the accuracy of the structural projections and their viability

Evaluation of the Methodology

Simple Typological Discovery

Typological Parameters

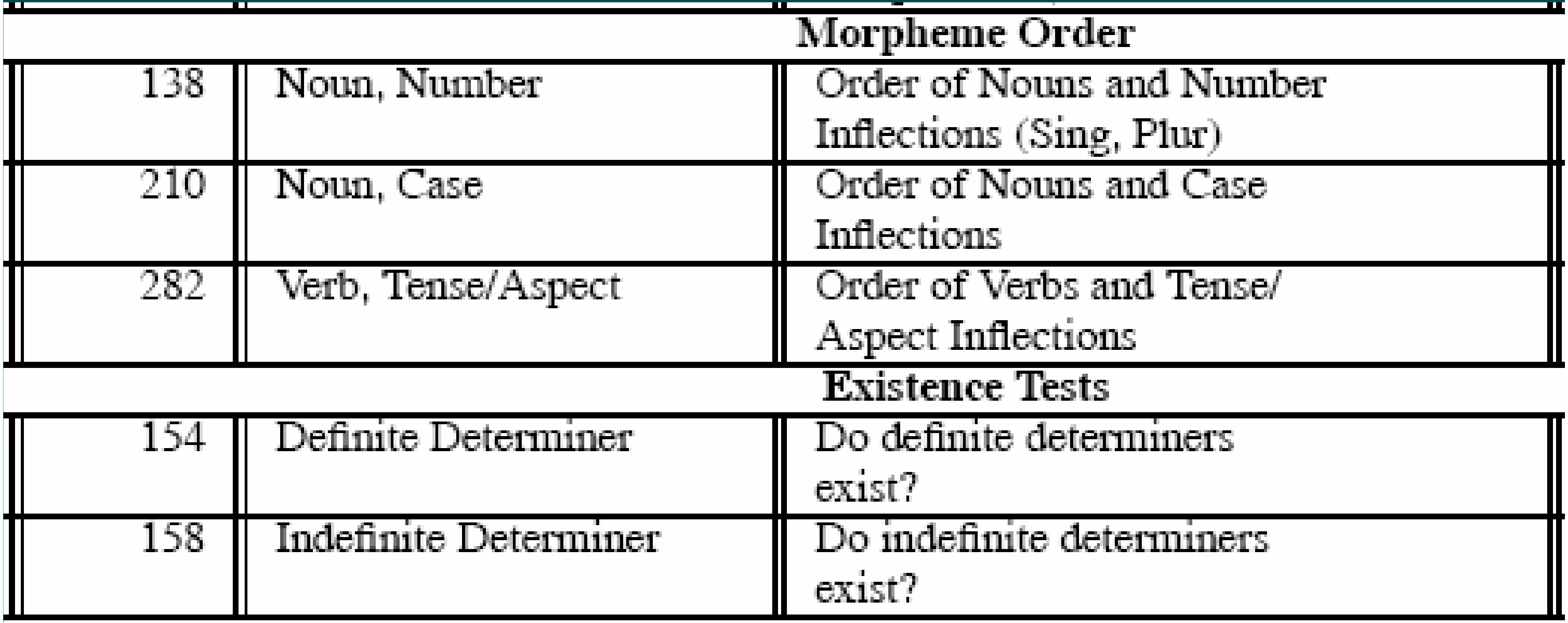
- From WALS (Haspelmath et al 2005)



WALS #	**parameter**	**Description**
Word Order		
330	Sentential Word Order	Order of Words in a sentence
342	Order of Verb and Objects	Order of the Verb, Object and and Oblique Object (e.g., PP) in the VP
N/A	Definite/Indefinite Determiners, Noun	Order of Nouns and Determiners *a*, *the*
358	Demonstrative, Noun	Order of Nouns and Demonstrative Determiners (*this*, *that*)
354	Adjective, Noun	Order of Adjectives and Nouns
N/A	Possessive Pronoun, Noun	Order of Possessive Pronouns and Nouns
350	Possessive NP, Noun	Order of a Possessive NPs and Nouns
346	Adposition, Noun	Order of Adpositions (*e.g.*, Preposition, Postpositions) and Nouns

Typological Parameters

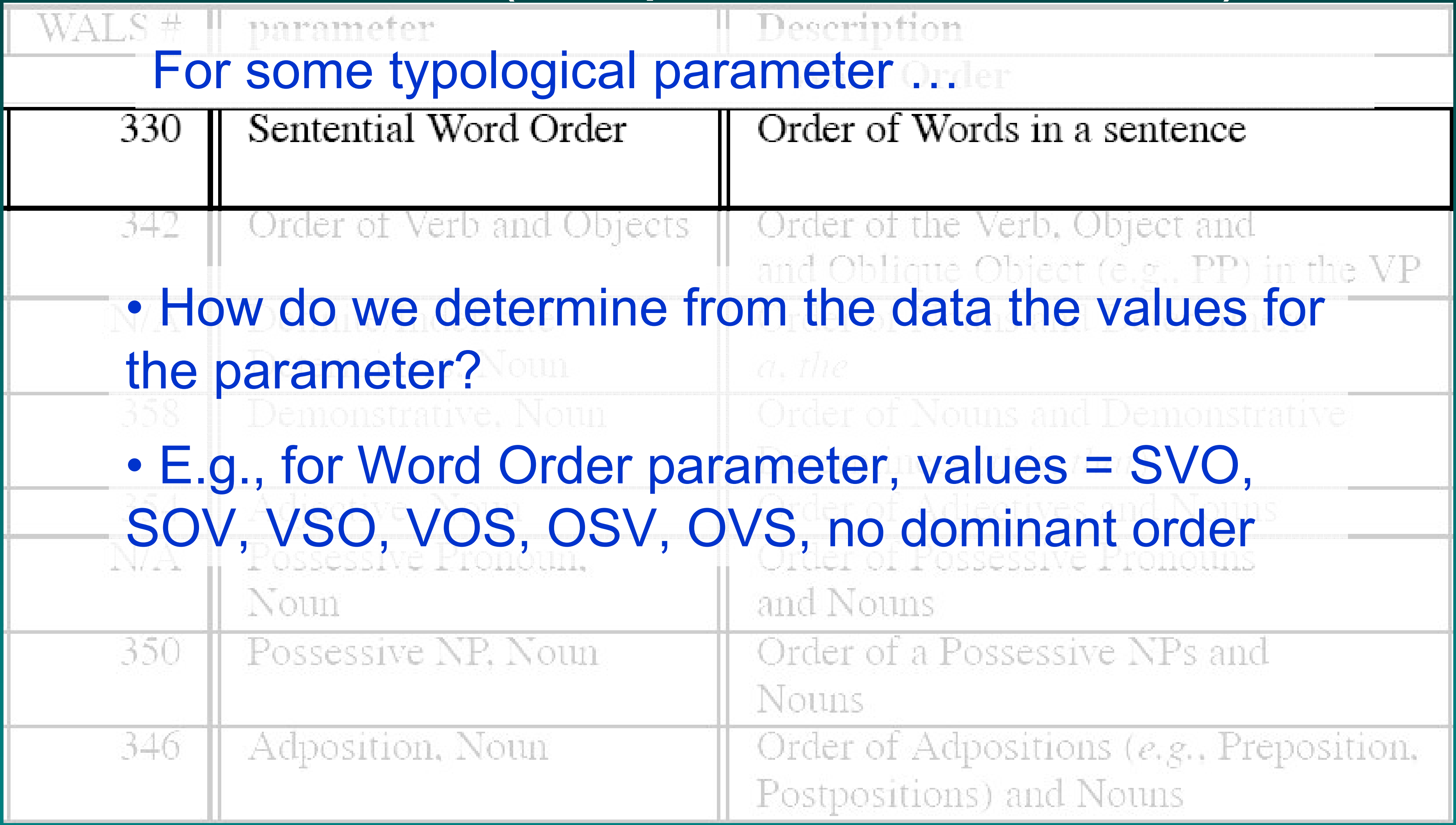
- From WALS (Haspelmath et al 2005)



Morpheme Order		
138	Noun, Number	Order of Nouns and Number Inflections (Sing, Plur)
210	Noun, Case	Order of Nouns and Case Inflections
282	Verb, Tense/Aspect	Order of Verbs and Tense/ Aspect Inflections
Existence Tests		
154	Definite Determiner	Do definite determiners exist?
158	Indefinite Determiner	Do indefinite determiners exist?

Typological Parameters

- From WALS (Haspelmath et al 2005)



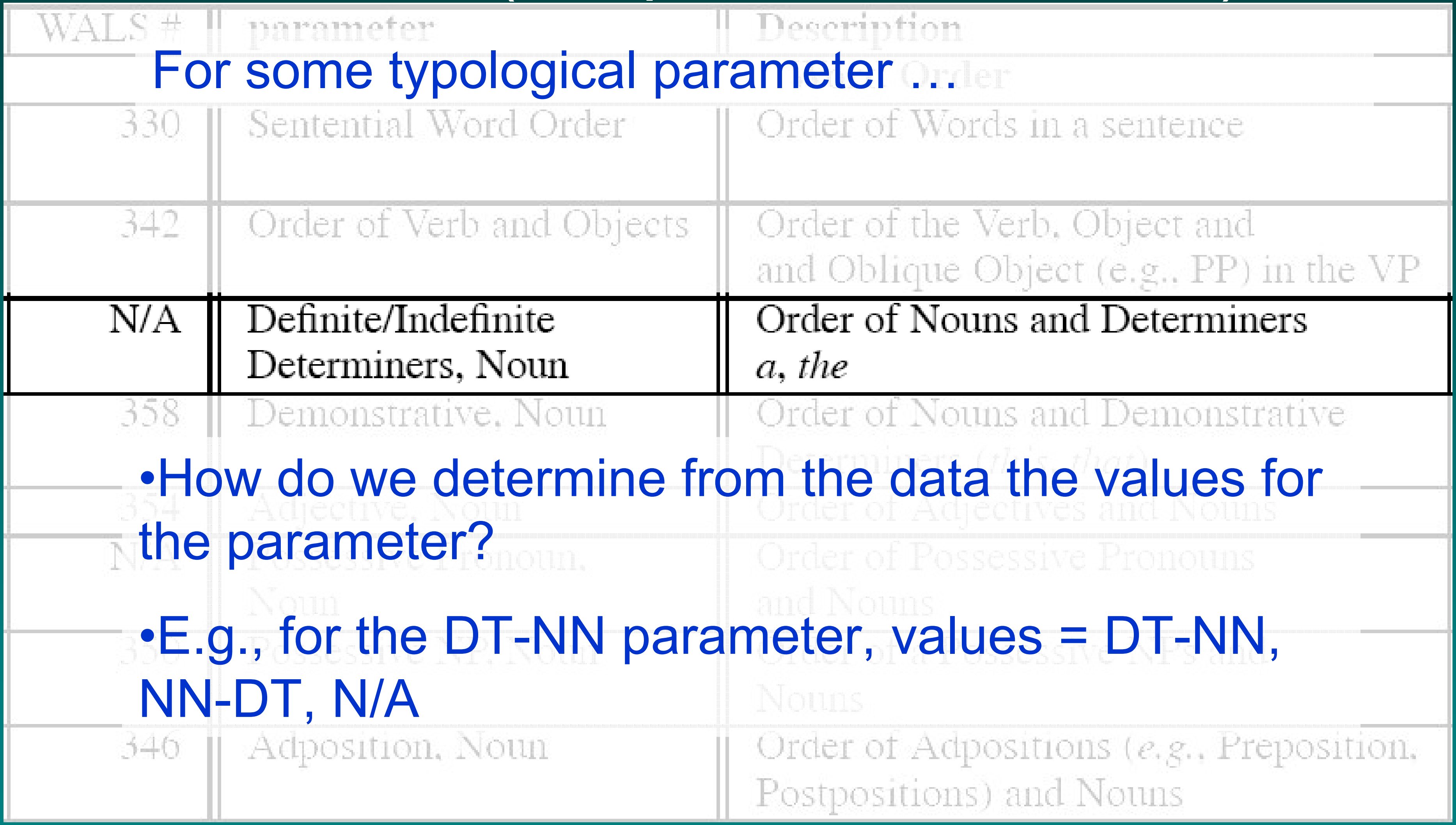
WALS #	parameter	Description
[illegible]	[illegible]	[illegible] Order
330	Sentential Word Order	Order of Words in a sentence
342	Order of Verb and Objects	Order of the Verb, Object and and Oblique Object (e.g., PP) in the VP
N/A	[illegible] Noun	Order of [illegible] *a, the*
358	Demonstrative, Noun	Order of Nouns and Demonstrative [illegible]
354	Adjective, Noun	Order of Adjectives and Nouns
N/A	Possessive Pronoun, Noun	Order of Possessive Pronouns and Nouns
350	Possessive NP, Noun	Order of a Possessive NPs and Nouns
346	Adposition, Noun	Order of Adpositions (*e.g.*, Preposition, Postpositions) and Nouns

For some typological parameter …

- How do we determine from the data the values for the parameter?
- E.g., for Word Order parameter, values = SVO, SOV, VSO, VOS, OSV, OVS, no dominant order

Typological Parameters

- From WALS (Haspelmath et al 2005)



WALS #	parameter	Description
[illegible]	[illegible]	Order
330	Sentential Word Order	Order of Words in a sentence
342	Order of Verb and Objects	Order of the Verb, Object and and Oblique Object (e.g., PP) in the VP
N/A	Definite/Indefinite Determiners, Noun	Order of Nouns and Determiners *a, the*
358	Demonstrative, Noun	Order of Nouns and Demonstrative Determiners (*this, that*)
354	Adjective, Noun	Order of Adjectives and Nouns
N/A	Possessive Pronoun, Noun	Order of Possessive Pronouns and Nouns
[illegible]	Possessive NP, Noun	Order of a Possessive NPs and Nouns
346	Adposition, Noun	Order of Adpositions (*e.g.*, Preposition, Postpositions) and Nouns

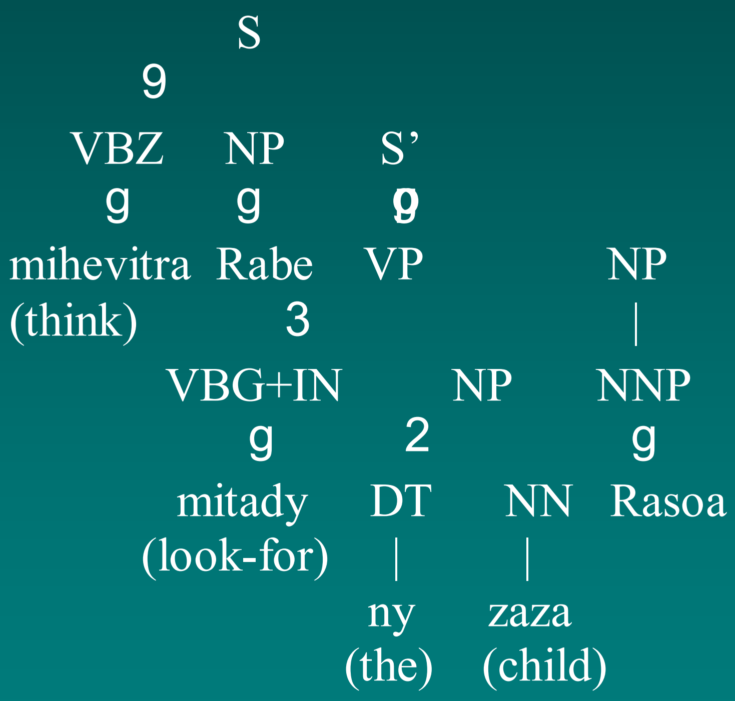
For some typological parameter …

- How do we determine from the data the values for the parameter?
- E.g., for the DT-NN parameter, values = DT-NN, NN-DT, N/A

Determining Value for a Typological Parameter

- Requires looking across sample of *annotated* data for language
- That is, a sample of the relevant Context Free Grammar (CFG) rules for the language
- Building CFGs from annotated data requires:
  - Distilling all trees for projected structures into grammar for the language
  - Collapsing identical rules and tabulating frequencies

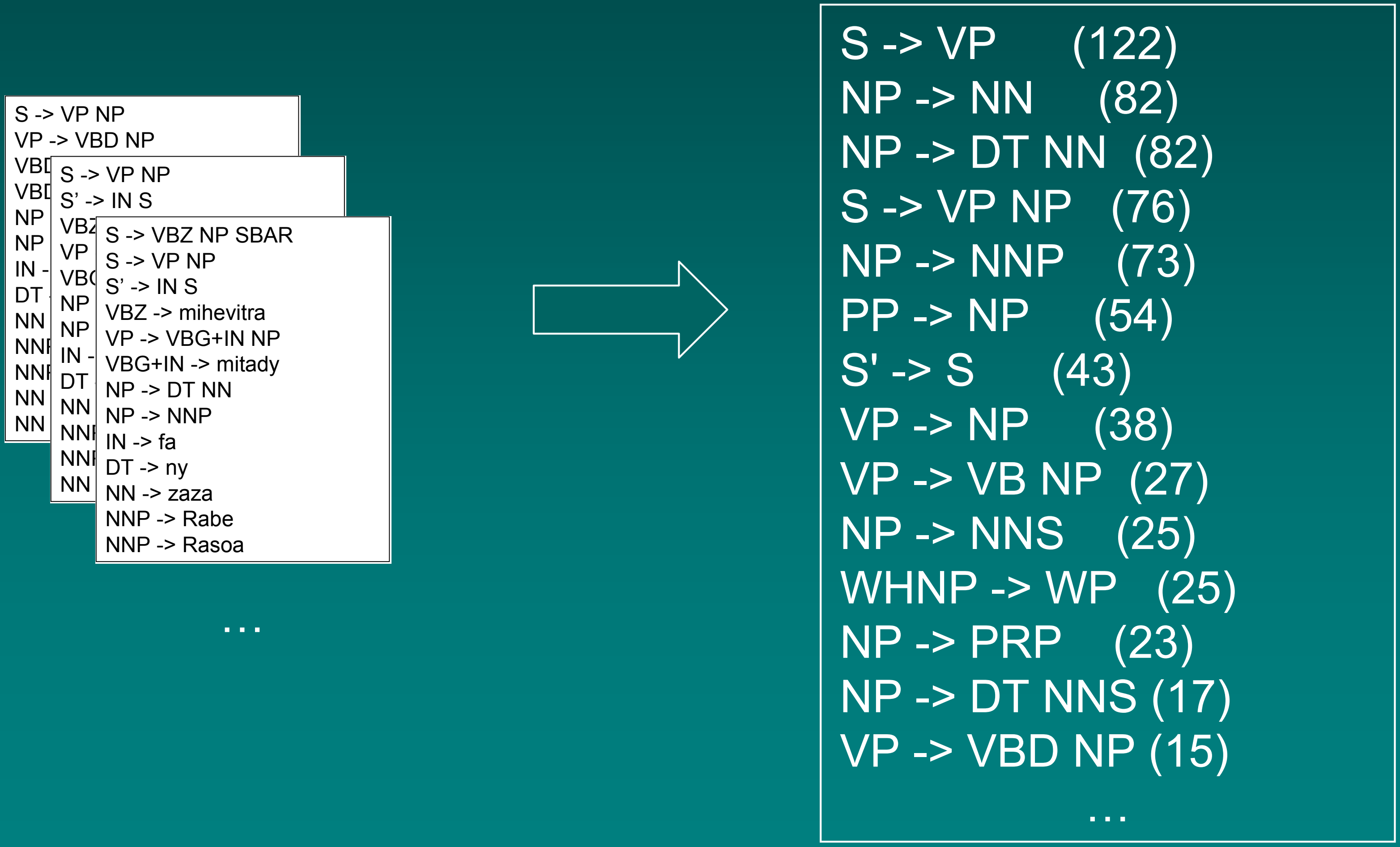
Distill Projected Trees into CFGs



S -> VBZ NP SBAR
S -> VP NP
S’ -> IN S
VBZ -> mihevitra
VP -> VBG+IN NP
VBG+IN -> mitady
NP -> DT NN
NP -> NNP
IN -> fa
DT -> ny
NN -> zaza
NNP -> Rabe
NNP -> Rasoa

Malagasy: Polinksy & Potsdam 2005

Collapse Identical Rules, Calculate Frequencies



S -> VP NP
VP -> VBD NP
VBD
VBD
NP
NP
IN -
DT
NN
NNP
NNP
NN
NN

S -> VP NP
S’ -> IN S
VBZ
VP
VBG
NP
NP
IN -
DT
NN
NNP
NNP
NN

S -> VBZ NP SBAR
S -> VP NP
S’ -> IN S
VBZ -> mihevitra
VP -> VBG+IN NP
VBG+IN -> mitady
NP -> DT NN
NP -> NNP
IN -> fa
DT -> ny
NN -> zaza
NNP -> Rabe
NNP -> Rasoa

…

S -> VP	(122)
NP -> NN	(82)
NP -> DT NN	(82)
S -> VP NP	(76)
NP -> NNP	(73)
PP -> NP	(54)
S' -> S	(43)
VP -> NP	(38)
VP -> VB NP	(27)
NP -> NNS	(25)
WHNP -> WP	(25)
NP -> PRP	(23)
NP -> DT NNS	(17)
VP -> VBD NP	(15)
…	

Determining Value for the Determiner-Noun Parameter

- For DT-NN, need NP rules

S -> VP	(122)
NP -> NN	(82)
NP -> DT NN	(82)
S -> VP NP	(76)
NP -> NNP	(73)
PP -> NP	(54)
S' -> S	(43)
VP -> NP	(38)
VP -> VB NP	(27)
NP -> NNS	(25)
WHNP -> WP	(25)
NP -> PRP	(23)
NP -> DT NNS	(17)
VP -> VBD NP	(15)
…	

DT-NN language

Determining Value for the Word Order Parameter

- For Word Order Parameter, need S and VP rules (or linear order in S rule)
- Problem: Identity of NPs unclear
- Idea: functionally tag English, and project

S -> VP	(122)	
NP -> NN	(82)	
NP -> DT NN	(82)	
S -> VP NP	(76)	Vxx language?
NP -> NNP	(73)	
PP -> NP	(54)	
S' -> S	(43)	
VP -> NP	(38)	
VP -> VB NP	(27)	VOS? VSO?
NP -> NNS	(25)	
WHNP -> WP	(25)	
NP -> PRP	(23)	
NP -> DT NNS	(17)	
VP -> VBD NP	(15)	
…		

Additional Annotations

- NP-SUBJ, NP-OBJ – mark subjects and objects
- PP-XOBJ, NP-XOBJ – mark oblique objects
- NP-Poss – Possessive NP
- DT1-4 – Marks various kinds of determiners (definite, indefinite, deictic, all others)
- Many other annotations possible (e.g., semantic roles, construction specific tags, etc.)

Determining Value for the Word Order Parameter

- CFG with functional tags projected

S -> VP	(122)	
S -> VP NP-SBJ	(64)	Vxx language
NP-SBJ -> NNP	(54)	
S' -> S	(43)	
PP-XOBJ -> NP	(38)	
NP-SBJ -> DT NN	(38)	
NP-OBJ -> NN	(37)	
NP -> NN	(36)	
VP -> NP-OBJ	(34)	
VP -> VB NP-OBJ	(25)	VOS!
WHNP -> WP	(25)	
NP-OBJ -> DT NN	(24)	
NP -> DT NN	(19)	
…		

Experiments 1&2

- For 10 languages
  - Determine values for 14 parameters
  - Evaluate against WALS (12) or other sources (2)
- Experiment 1
  - Use no functional tags (only phrasal & POS)
- Experiment 2
  - Use functional tags (e.g., NP-SUBJ, etc.)

Results

Parameter	CFG	CFG+func
WOrder	80%	90%
VP-OBJ	50%	60%
DT-NN	80%	80%
Dem-NN	80%	90%
JJ-NN	100%	100%
PRP$-NN	80%	80%
Poss-NN	60%	70%
P-NP	90%	90%
number	70%	70%
case	80%	80%
T/A	80%	80%
Def	100%	100%
Indef	90%	90%
Mean	80%	83%

Experiment 3

- Project Structures for 98 languages
- Determine value of WOrder parameter for each language (e.g., SVO, SOV, etc.)
  - How much data is required for accurate answers?
  - What's the relationship between the number of IGT examples and the probability of a correct answer?

Results

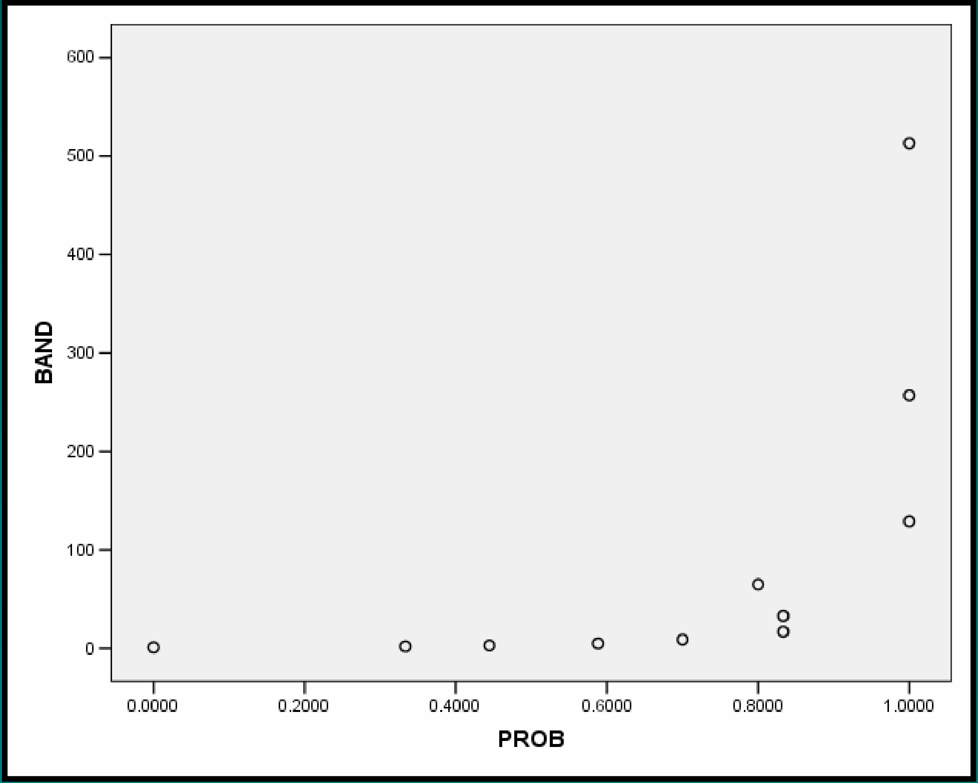
- Accuracy: For 69 of the 98 languages, WOrder was accurately determined
- Confusion matrix:

		Guess			
		SVO	SOV	VSO	VOS
Actual	SVO	32	8	0	9
	SOV	2	33	0	6
	VSO	2	2	3	4
	VOS	0	0	0	1

Results

- Accuracy improved as # of IGT instances increased

# IGT	Avg. Accuracy
100+	100%
40-99	99%
10-39	79%
5-9	65%
3-4	44%
1-2	14%



What the Results Show

- We can fairly accurately discern values for several typological parameters
  - English bias of projections has minimal effects (on these parameters)
- Larger samples overcome the effects of
  - IGT Bias
- We can do this across data for many languages *automatically*
- Might generalize to some other parameters
- We can return data
- See Lewis & Xia 2008 (IJCNLP) for more details

Summary and Future Work

Summary

- We demonstrate
  - A tool that was built automatically from language data found on the Web
  - ML techniques (detection, lang ID) that improve both precision and recall
  - The potential for resources composed of 100s of languages and 1000s of data points for automated analysis and discovery
  - How to work within Copyright Law and linguistics custom when serving up data

Future Directions

- Using ML techniques, scale up ODIN's size
- Improve query infrastructure
  - Support richer query across language data
  - Support freer-form user queries (tgrep2)
- Building deep grammars
  - Seed Bender's Matrix project (HPSG) (Bender et al 2002)
    - Answer typological queries + provide data from ODIN
    - Create seeds for building deep grammar fragments
- Create transfer rules for MT work (Fox 2002)
- Evaluate structural divergence on scale (Xia and Lewis, under revision)
- Bootstrap tool development (Lewis 2006)

Project Specific References

Overview:

- Lewis, William and Fei Xia (2009). 'Parsing, Projecting & Prototypes: Repurposing Linguistic Data on the Web', in *Proceedings of the European Association of Computational Linguistics (EACL) Conference*, Athens, Greece, March 2009.
- Lewis, William (2006), 'ODIN: A Model for Adapting and Enriching Legacy Infrastructure', in *Proceedings of the e-Humanities Workshop, held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam.

Typological Discovery:

- Lewis, William and Fei Xia (2008). 'Automatically Identifying Computationally Relevant Typological Features', in *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, January 2008.

Projection:

- Xia, Fei and William Lewis (2007), 'Multilingual Structural Projection across Interlinearized Text', in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)/HLT*, Boston, April 2007.
- Lewis, Xia, and Jinguji (2006). 'Enriching Language Data through Projected Structures', *Proceedings of the Texas Linguistics Conferences 10 (TLSX)*, Austin, Texas, October.

Language ID:

- Xia, Fei, William Lewis, and Hoifung Poon (2009). 'Language ID in the Context of Harvesting Language Data off the Web', in *Proceedings of the European Association of Computational Linguistics (EACL) Conference*, Athens, Greece, March 2009.

IGT Detection:

- Xia and Lewis (2008). 'Repurposing Theoretical Linguistic Data for Tool Development and Search', in *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, January 2008.

Infrastructure:

- Farrar, Scott and William Lewis (2006). *The GOLD Community of Practice: An Infrastructure for Linguistic Data on the Web*. Journal of Language Resources and Evaluation.

ODIN: http://www.csufresno.edu/odin LinguistList: http://odin.linguistlist.org

References

- Bailyn, John F. 2001. “Inversion, Dislocation and Optionality in Russian.” In Zybatow, Gerhild. *Current Issues in Formal Slavic Linguistics*.
- Bender, Emily M., Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Taipei, Taiwan.
- Charniak, Charniak. 1997. “Statistical Parsing with a Context-Free Grammar and Word Statistics.” In *Proceedings of AAAI-1997*.
- Fox, Heidi. 2002. Phrasal cohesion and statistical machine translation. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing.*
- Haspelmath, Martin, Mathew Dryer, David Gil, and Bernard Comrie. 2005. *World Atlas of Language Structures.* Oxford University Press.
- Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. “Evaluating translational correspondence using annotation projection.” In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, Pennsylvania.
- Martínez-Fabián, Constantino. 2006. *Yaqui Coordination*. Unpublished dissertation. University of Arizona.
- Polinsky, Maria and Eric Potsdam. 2005. Malagasy Control and Its Theoretical Implications. *Proceedings of the Berkeley Linguistic Society Annual Meeting*.
- Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. “Dependency tree translation: Syntactically informed phrasal SMT.” In *Proceedings of ACL 2005*.
- Xi, Chenhai and Rebecca Hwa. 2005. “A backoff model for bootstrapping resources for non-English languages.” In *Proceedings of HLT-EMNLP*, pages 851–858, Vancouver, British Columbia, Canada.
- Yarowksy, David and Grace Ngai. 2001. “Inducing multilingual pos taggers and NP bracketers via robust projection across aligned corpora.” In *Proceedings of NAACL-2001*, pages 377–404.