

Building NLP Systems to Assist Linguistic Studies

Fei Xia

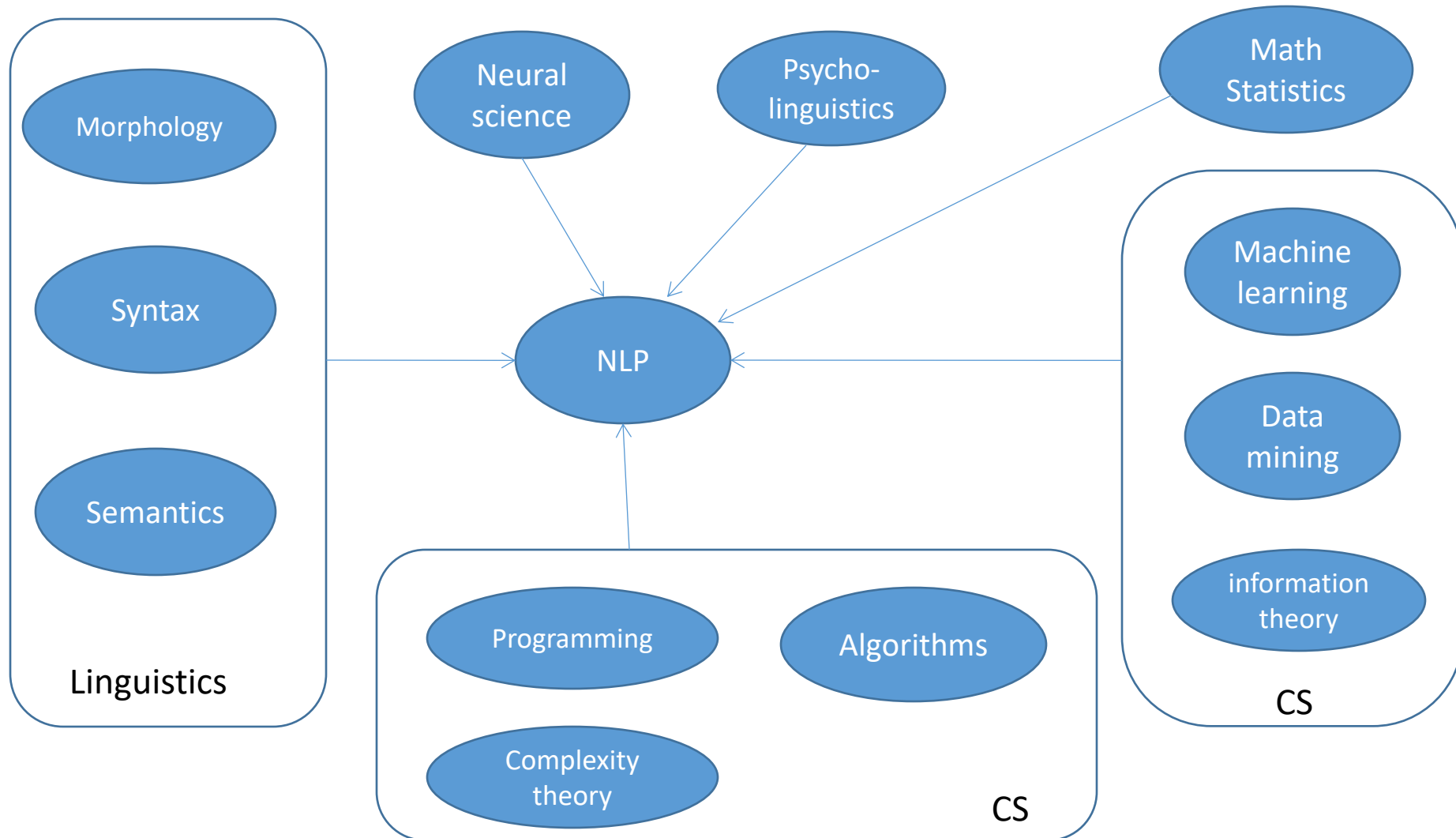
Linguistics Department
University of Washington

fxia@uw.edu

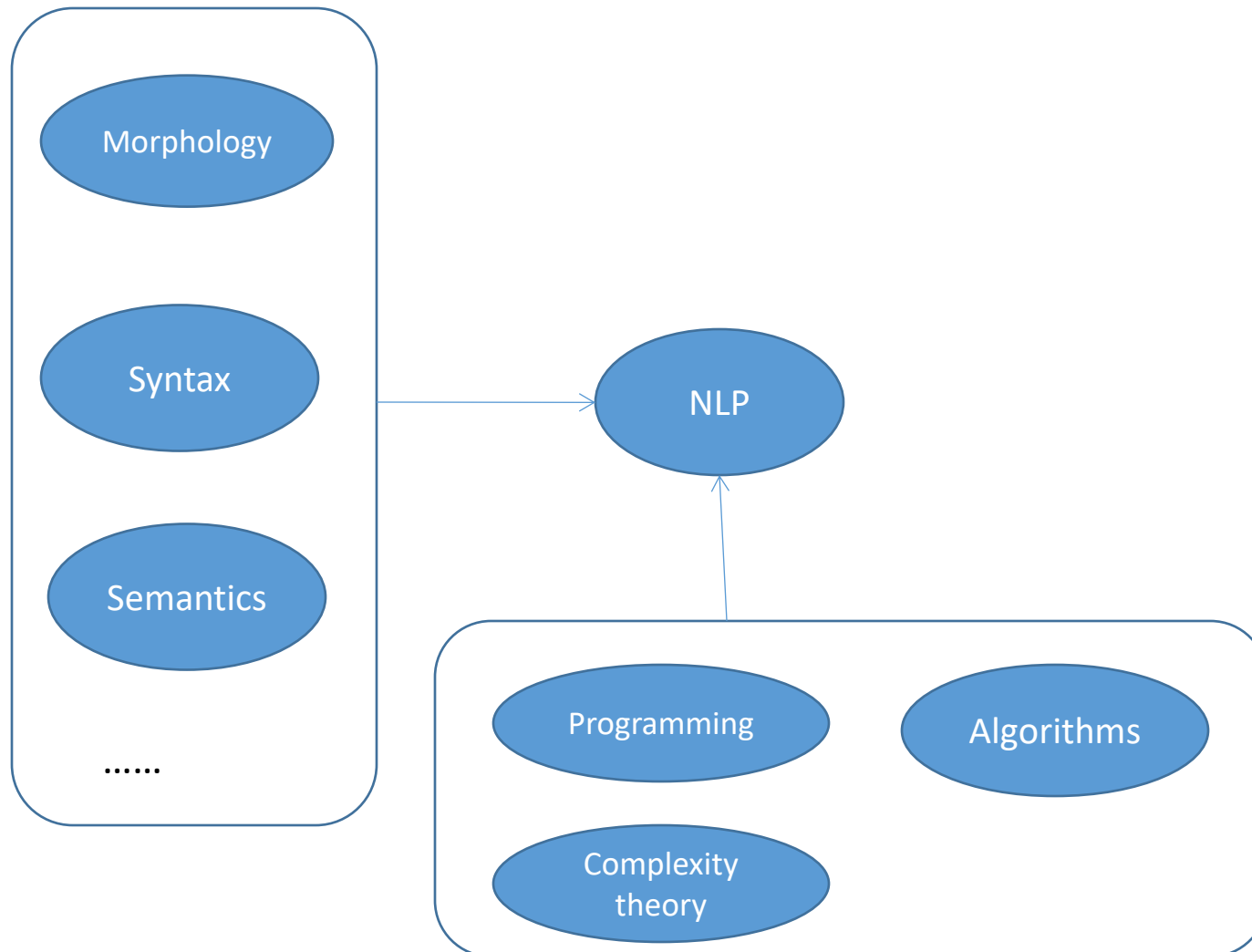
Outline

- Natural language processing (NLP) and linguistics
- The RiPLes project: collecting and enriching IGT data
 - Joint work with William Lewis and Ryan Georgi
- The AGGREGATION project: Inferring grammars from the enriched IGT
 - Joint work with Emily Bender, Michael Goodman, Olga Zamaraeva, and Kristen Howell

NLP and related fields

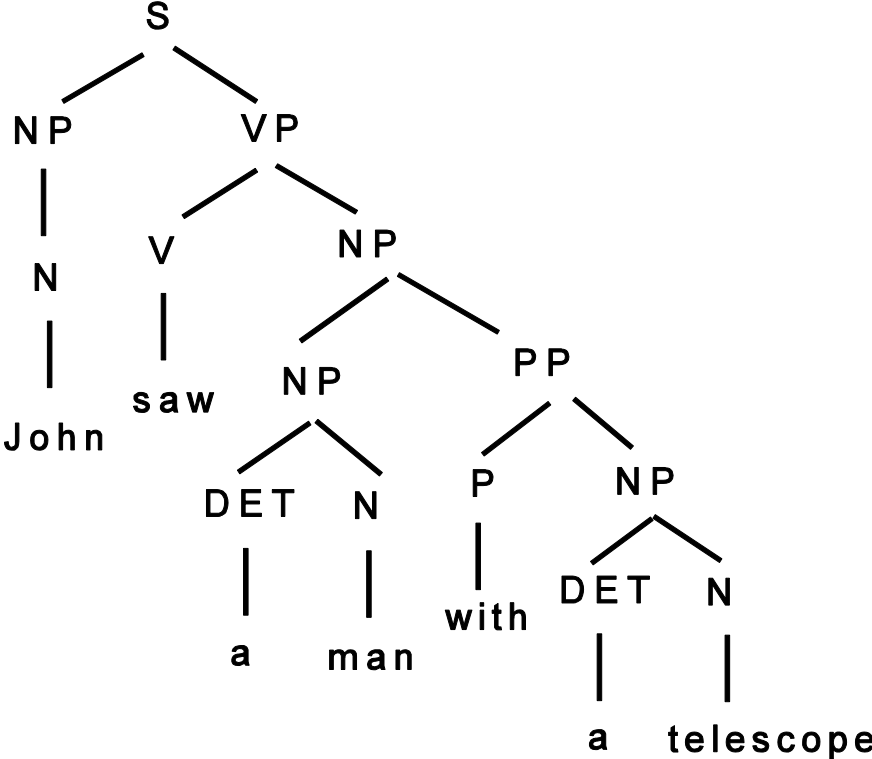
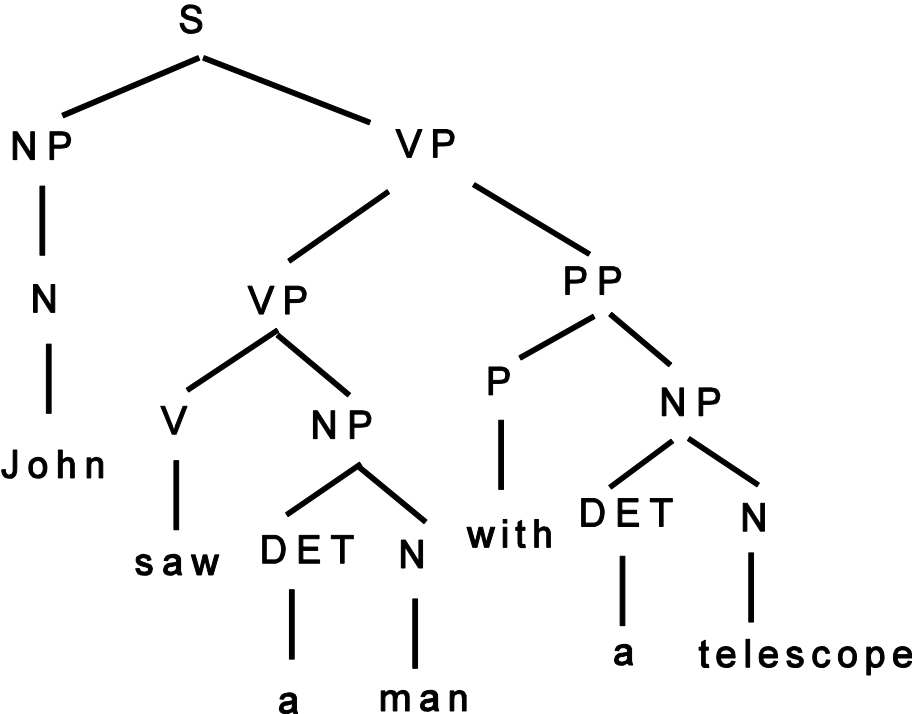


Early NLP: rule-based approach



Parsing

John saw a man with a telescope

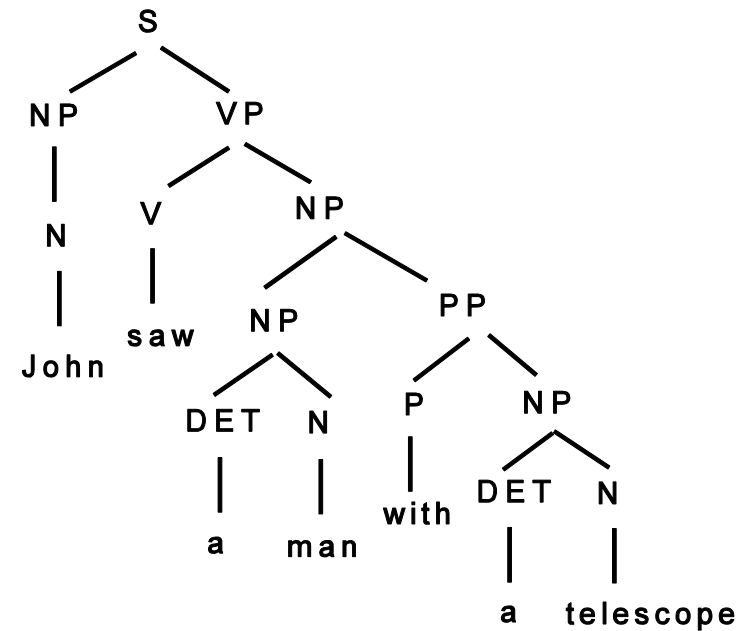
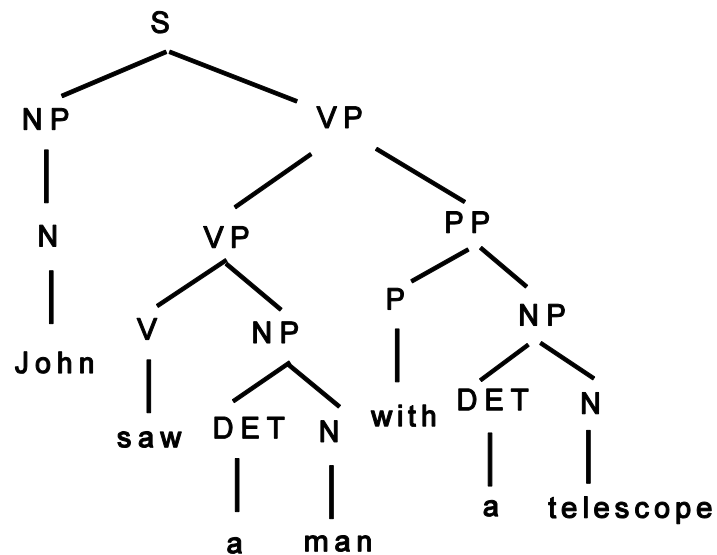


Syntax vs. Parsing

Syntax: It does not have any deep feeling.

*It has any deep feeling.

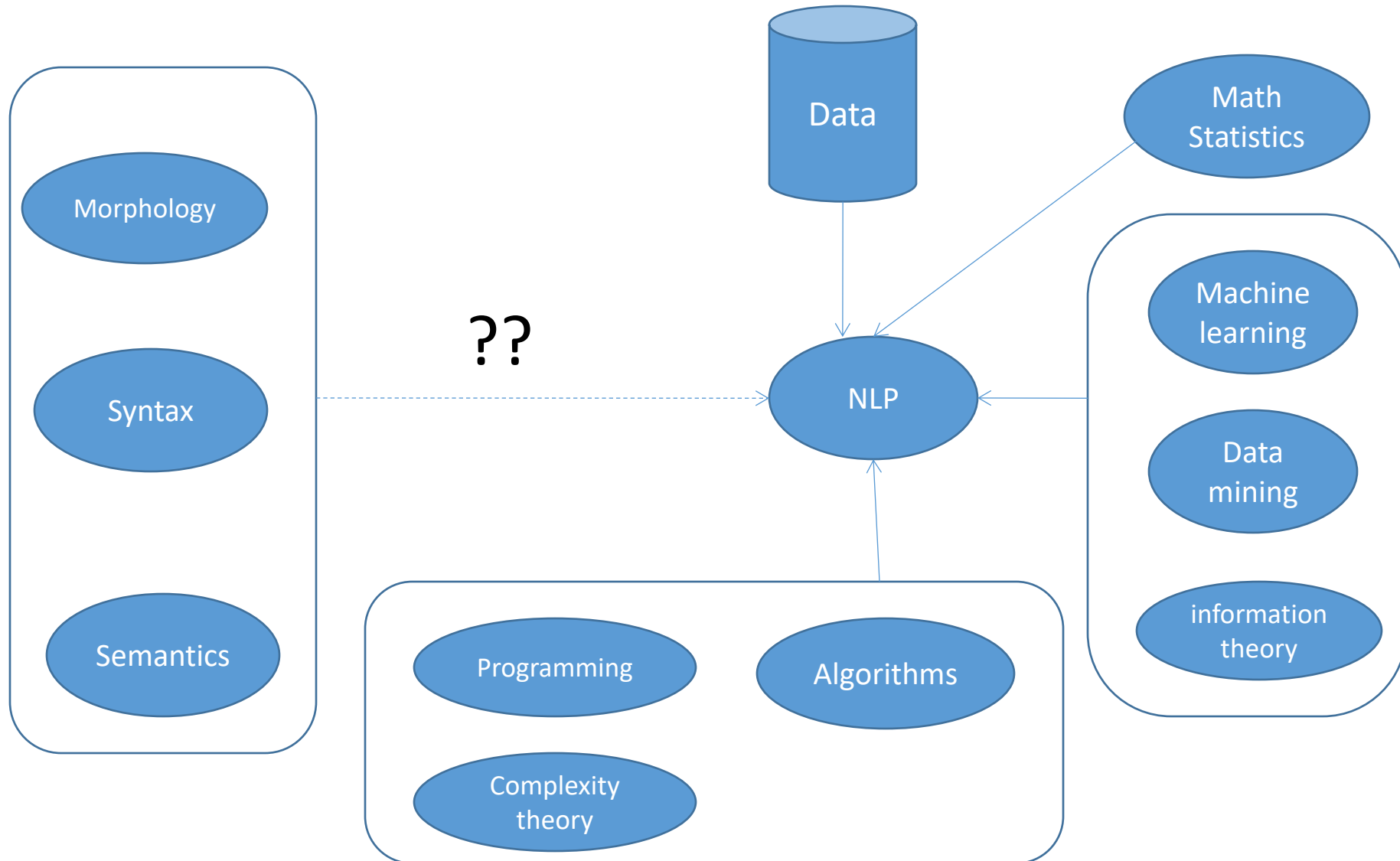
Parsing: John saw a man with a telescope



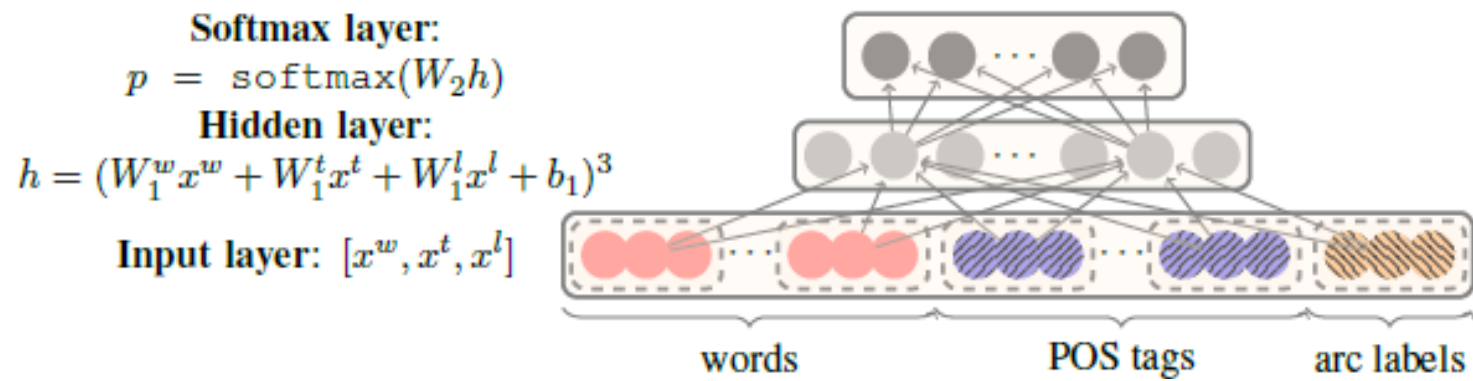
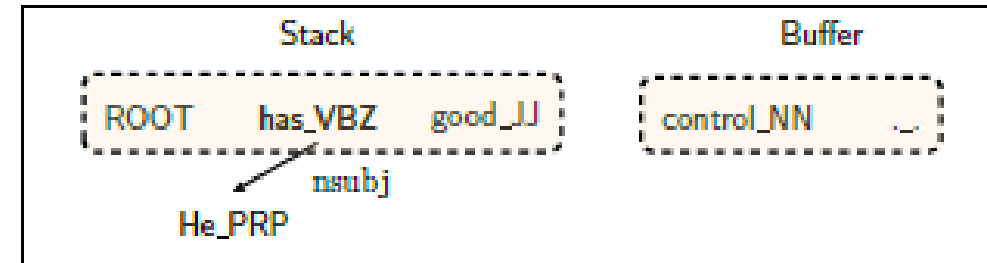
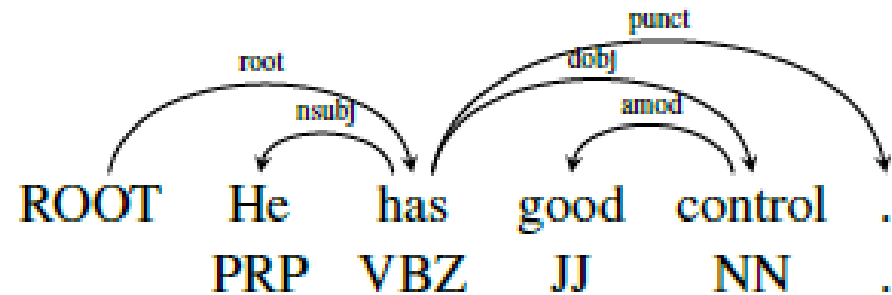
Parsing: Rule-based approach

- Steps:
 - Choose a framework: e.g., CFG, LTAG, HPSG, LFG, CCG
 - Build a grammar in that framework
 - Use the grammar to parse a test suite and check the parse trees
 - Revise the grammar to improve coverage and reduce ambiguity
- Advantages:
 - Intuitive and linguistically motivated: $S \Rightarrow NP VP$
 - Can capture generalization: e.g., agreement between subject and verb
- Disadvantages:
 - Require linguistic expertise: e.g., build a grammar for Arabic/Urdu/Farsi/Chintang
 - Long development cycle: many person years
 - Not robust:
 - “Ungrammatical” sentences: repair, errors, corruption
 - Unknown words, new constructions, novel usage
 - Rules are not good at handling ambiguity: John saw a man with a telescope

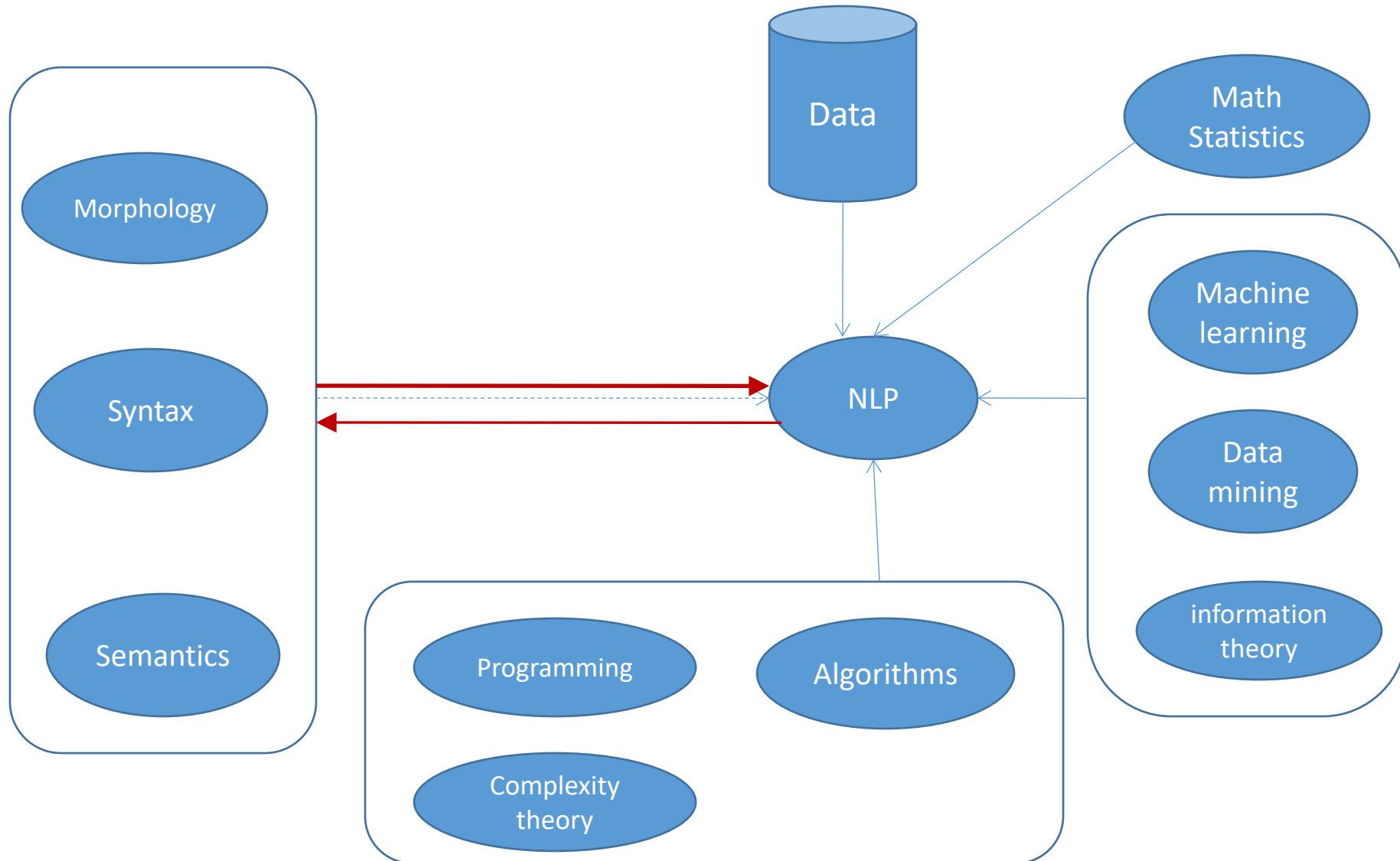
Current NLP: statistical/neural approach



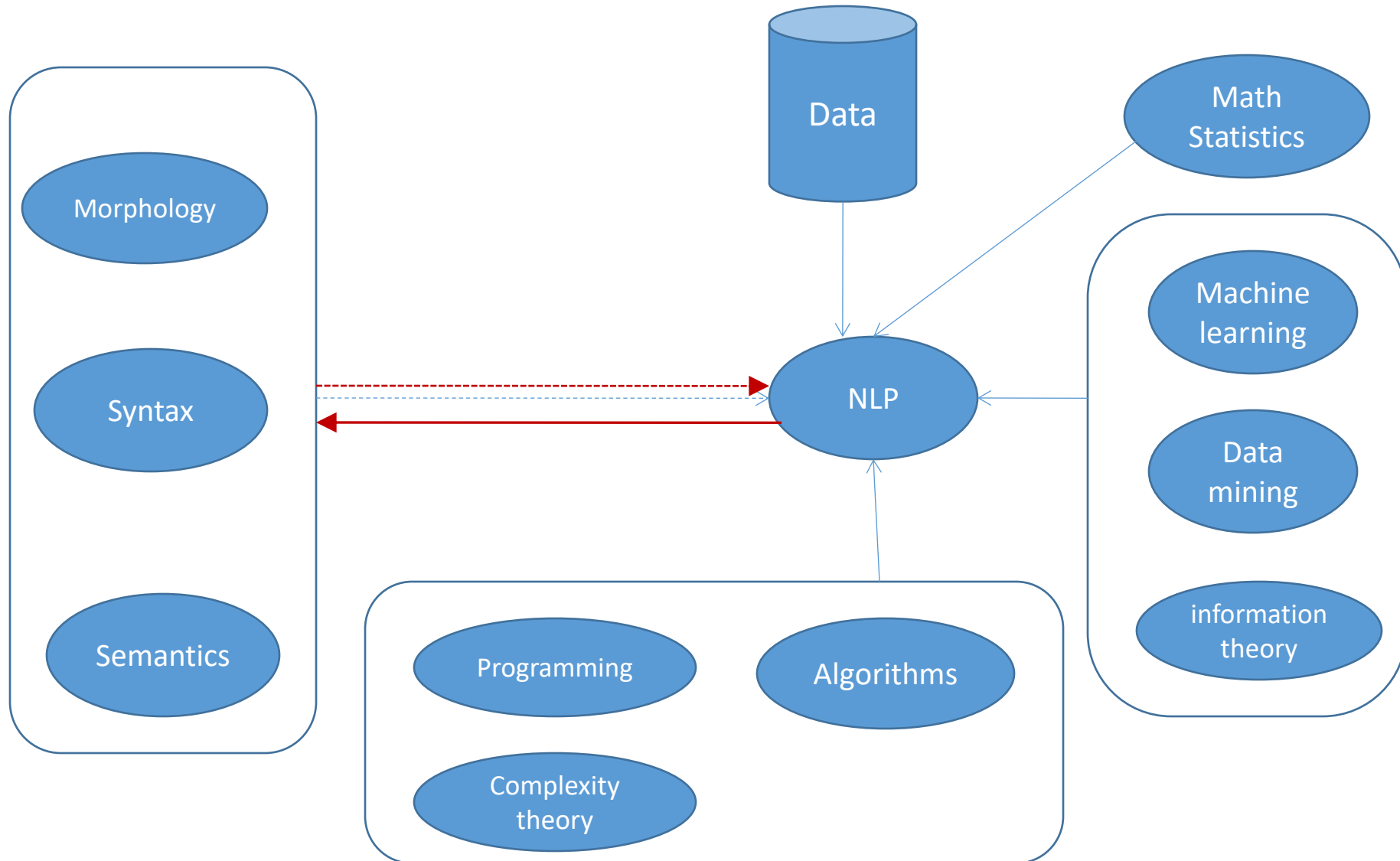
Parsing using Neural Networks: an example from (Chen and Manning, 2014)



My research interest



My research interest



Outline

- Natural Language Processing (NLP) and linguistics
- **The RiPLes project**
 - Motivation
 - ODIN: Collecting language data from the Web
 - INTENT: Enrich the data
- The AGGREGATION project

Low-resource languages

- There are more than 7000 living languages in the world.
- 90% of them are likely to go extinct or become seriously threatened in the next 100 years (Krauss, 1992).
- Most languages are low resource: some with millions of speakers, while others with only a few native speakers left.
- We know very little about most of these languages.

In the linguistics field

- Ethnologue covers more than 7,000 living languages (Gordon, 2005): info under several categories (e.g., countries, population, language code)
- The World Atlas of Language Structures (WALS) : a large database of structural properties of languages
 - 144+ features: e.g., what is the word order of a language
 - 2670+ languages, 55+ authors
 - Among all possible (feat, lang) pairs, about 15% of them are specified by the authors.

Motivation: Linguistics

- Questions:
 - For a particular language:
 - ❖ What is word order: SVO, SOV, VSO,?
 - ❖ Does this language allow long-distance scrambling?
 - For cross-language study:
 - ❖ Find all the languages that allow long-distance scrambling and examples in those langs
 - ❖ Implicational universal: **if determiners follow nouns, then relative clauses will also follow nouns**. Is it always true? If not, find exceptions.
- How can we automate the process of locating data and answering these questions?

Motivation: NLP

- Goal: to create tools (e.g., parsers) for many languages
- Most common approach: supervised learning
 - Problem: Creating **labeled** data (e.g., treebanks) can be very expensive.
 - Solution: unsupervised learning with **unlabeled** data, transfer learning
- A small amount of linguistic knowledge often helps a lot:
 - Ex: Prototypes (e.g., **NP** → **Det N**) for grammar induction: 26.3% to 65.1% for English (Haghighi and Klein, 2006):
- Question: How can we acquire the knowledge automatically for hundreds of languages?

In the NLP field

- Idea: Take advantage of existing resources for resource-rich languages
- Common approaches:
 - Use bitext and syntactic projection:
 - Use word alignment and then project information from one language to another
 - Issue: Many resource-poor languages do not have a lot of bitext.
 - ➔ We use Interlinear Glossed Text (IGT), and study structural divergence.
 - More recently, many studies on transfer learning with neural network.

Interlinear glossed text (IGT)

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
(Welsh, from (Bailyn 2001))

→ ODIN is a collection of IGT
(Online Database of Interlinear glossed text)

The RiPLes Project

- RiPLes stands for “information engineering and synthesis for Resource-Poor Languages”
- Main components:
 - ODIN: extract IGT data from linguistic documents
 - INTENT: enrich IGT data and use it to bootstrap NLP systems

Building ODIN (Lewis and Xia, 2010)

- Crawling the Web for documents that contain IGT
- IGT detection: locate IGT within the document
- Language ID: determine the language name/code for each IGT
- Manual check: final check before the release of ODIN

A linguistic document

Reconsidering structural case in Finnish

Dieter Wunderlich

This paper is a response to Kiparsky (2000), who convincingly argues that the complex case marking in Finnish can ...

..... The genitive is also blocked in the imperative construction, shown in (1b), as well as in the passive....

- (1) a. Tuo-n häne-t / karhu-n / karhu-t / karhu-j-a.
bring-1 sg (s)he-ACC / bear-GEN / bear-pl.NOM / bear-pl-PART
'I'll bring him/her / a/the bear/ the bears / bears'
- b. Tuo häne-t / karhu / karhu-t / karhu-j-a.
bring (s)he-ACC / bear-NOM / bear-pl.NOM / bear-pl-PART
'Bring him/her / a/the bear / the bears / bears!'

On the basis of the fact that bounded

A linguistic document

Reconsidering structural case in **Finnish**

Dieter Wunderlich

This paper is a response to Kiparsky (2000), who convincingly argues that the complex case marking in **Finnish** can ...

..... The **genitive** is also blocked in the **imperative construction**, shown in (1b), as well as in the **passive**....

- (1) a. Tuo-n häne-t / karhu-n / karhu-t / karhu-j-a.
bring-1 sg (s)he-ACC / bear-GEN / bear-pl.NOM / bear-pl-PART
'I'll bring him/her / a/the bear/ the bears / bears'
- b. Tuo häne-t / karhu / karhu-t / karhu-j-a.
bring (s)he-ACC / bear-NOM / bear-pl.NOM / bear-pl-PART
'Bring him/her / a/the bear / the bears / bears!'

On the basis of the fact that bounded

Crawling the Web

- Intuition: documents that contain the following tend to contain IGT:
 - Grams: e.g., -NOM (nominative) , -ACC (accusative)
 - Language names and language codes: e.g., Finnish, Malagasy
 - Drawn from the Ethnologue database (Gordon, 2005)
 - Linguists' names and the languages that they work on: e.g., Kiparsky
 - Drawn from the Linguist List's linguist database (linguistlist.org)
- Try different combinations of terms from these categories:
 - Ex: NOM+ACC+Finnish

IGT detection

- 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).'

...

IGT detection: rule-based approach

- (1) a. Tuo-n häne-t karhu-n / karhu-t / karhu-j-a.
bring-1 sg (s)he-ACC bear-GEN / bear-pl.NOM / bear-pl-PART
'I'll bring him/her a/the bear/ the bears / bears'

Regular expression:

```
\s* \(\d+\) .*
```

```
.+
```

```
\s* [\`|\`|\"] .+ [\'|\"] .*
```

Difficulty in IGT detection

[DP [DO Ku] [AGRP [Adj ketaran] AGRO [NP namwu]]]

a.

the big tree

(Kim, 1997)

- Two-part IGT
- Extra annotations and structure
- Pdf-to-txt conversion noise
- ...

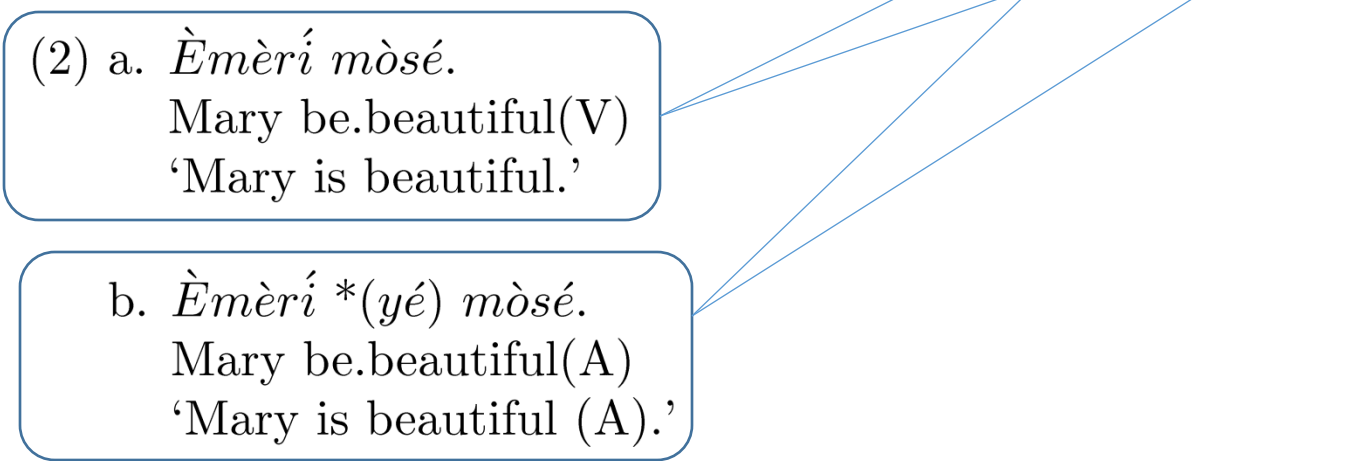
Applying NLP techniques to IGT detection

- Treat IGT detection as a sequence labeling problem.
- Label each line in a document with a BIO label:
 - B: the 1st line in an IGT
 - I: inside an IGT
 - E: the last line in an IGT
 - O: outside IGT
- Convert a tag sequence into IGT sequences by simple heuristics:
 - Ex: Any “B I* E” sequence is treated as an IGT instance.

Experiments

- Cues (aka “features”):
 - tokens that appear on the current line: e.g., -ACC
 - whether the current line starts an example number
 - whether the current line starts with a quotation mark
 - whether the current line and the previous line have similar indentation
 - ...
- Training data: 41 files with 1573 IGT
- Test data: 10 files with 447 IGT
- Results:
 - Exact match: 88.38% vs. 51.40% (RegEx)
 - Partial match: 95.40% vs. 74.58% (RegEx)

Language ID (Xia et al, 2009)

- 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).'
 - ...
- 
- The diagram consists of two blue rounded rectangular boxes. The first box contains lines 30-32, and the second box contains lines 34-36. Three blue arrows originate from the right side of these boxes and point towards the words 'Edo' and 'Kwa' in line 27.

Differences from a typical language ID task

- Large number of languages: 1000+
 - Unseen languages: 10% of IGTs in the test data belong to unseen languages
 - Very limited amount of training data: no more than 10 words per language for 45.3% of languages
 -
- TextCat (Cavnar and Trenkle's algorithm):
99.8% on 8 languages => 51.4% on ODIN data

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: Tiwa (Sino Tibetan) and Tiwa (Tanoan)
 - Ex: “Macrolanguages”: e.g., Chinese, Quechua
- We use language codes, because each language code maps to **exactly** one language.
- Our system will output both language code and language name.

Language tables

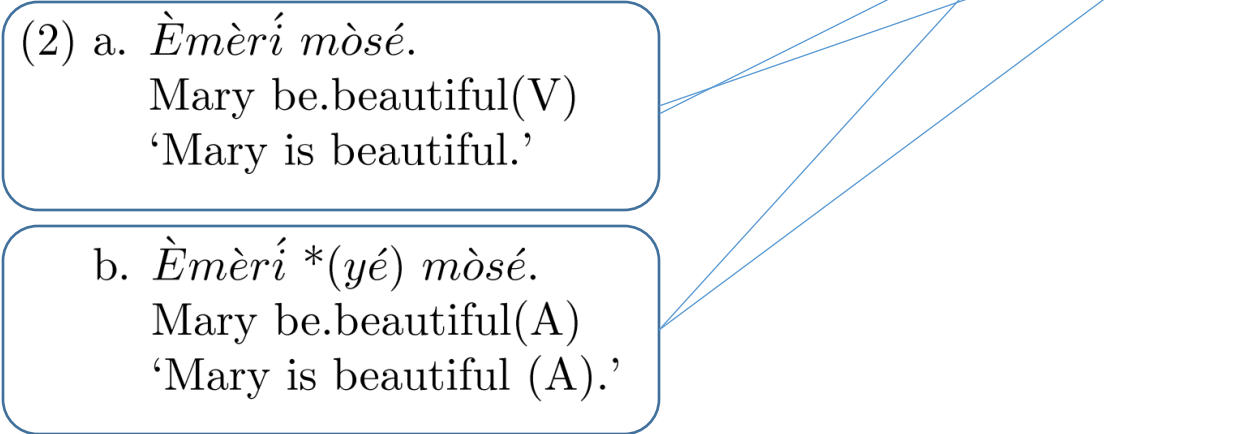
Language table	# of lang codes	# of lang (code, name) pairs
(1) ISO 639-3	7702	9312
(2) Ethnologue v15	7299	42789
(3) LinguistList table	231	232
Merged table	7816	47728

6% of language names in the merged table are ambiguous.

The table is not complete:

- Dozens of languages (e.g., Early High German) do not have language codes.
- More than 900 (code, name) pairs are missing from the table (e.g., Aroplokep vs. Arop-Lukep)

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).'
...
- 
- The diagram consists of two blue rounded rectangular boxes. The top box contains lines 30-32, and the bottom box contains lines 34-36. Three blue arrows originate from the right side of these boxes and point towards the citation lines 27 and 28, indicating that the examples are drawn from the source mentioned in those lines.

Treating language ID as a coreference task

- Coreference task:
 - Ex: **Bryan** called **Alisa**. **He** found **her** book.
 - A language name is like a proper noun.
 - An IGT is like a pronoun.
- Unseen languages are no longer a problem.
- All the existing algorithms on coreference in the NLP field can be applied to the task.

Experiments

- Features:
 - The language names that appear close to the current IGT
 - Word/character ngrams in the current IGT vs. ngrams for a language in the training data
 - Word/character ngrams in the current IGT vs. ngrams in other IGTs in the [same](#) document
- Data set: 1160 documents (90% training, 10% testing)
- Results:
 - 85.1% (CoRef) vs. 51.4% (TextCat)
 - with less training data: 81.2% (CoRef) vs. 28.9% (TextCat)

ODIN database (before manual correction)

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

The IGT is extracted from **3,000** documents.

ODIN database (after manual correction)

Range of IGT instances	# of languages	# of IGT instances	% of IGT instances
> 10000	1	10,814	6.88
1000-9999	31	81,218	51.69
100-999	139	46,420	29.55
10-99	460	15,650	9.96
1-9	862	3,012	1.92
Total	1,493	157,114	100

The IGT is extracted from 2025 documents.

ODIN v2.1 is at https://uakari2.ling.washington.edu/corpus/odin/v2_1/09384dc6/

Outline

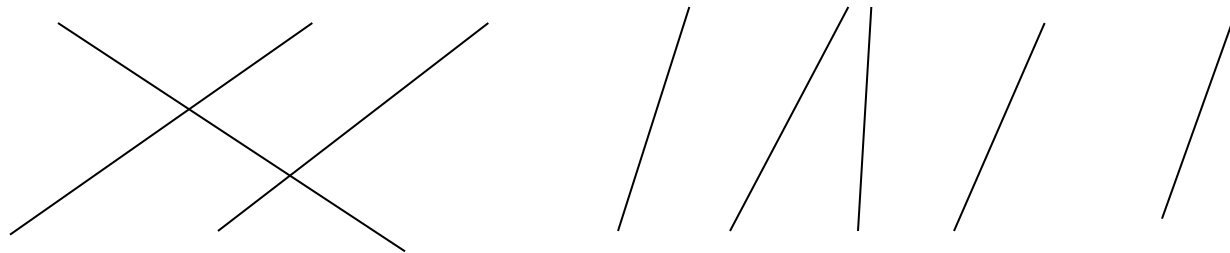
- Natural Language Processing (NLP) and linguistics
- The RiPLes project
 - Motivation
 - ODIN: Collecting language data from the Web
 - **INTENT: Enrich the IGT data**
- The AGGREGATION project

Enriching IGT: Align three lines

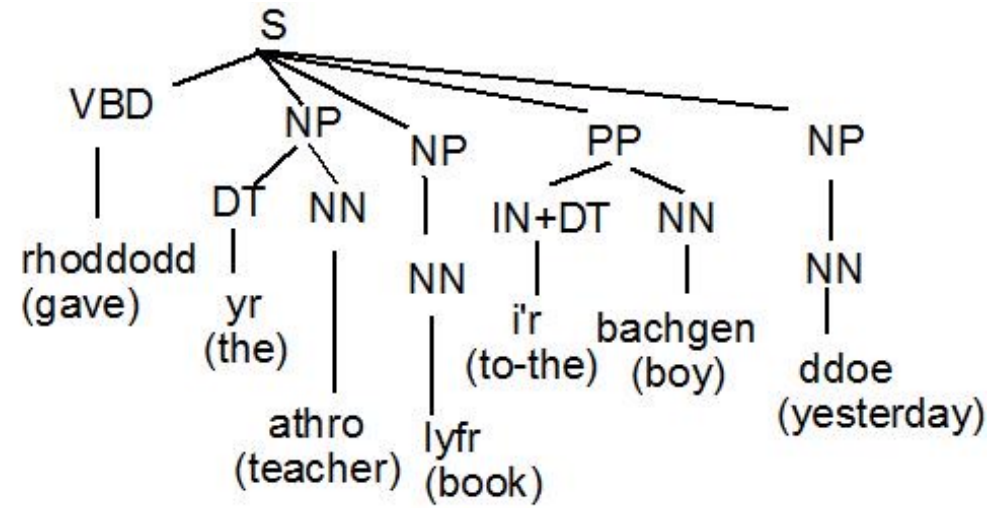
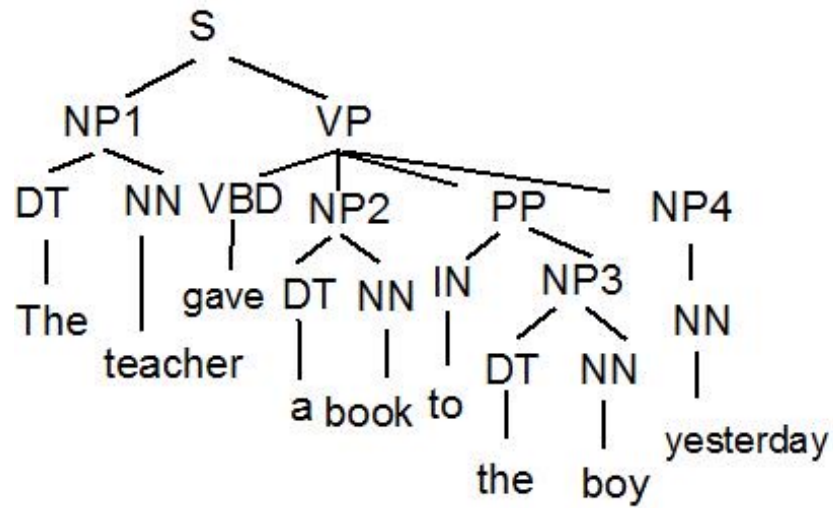
Rhoddodd yr athro lyfr i'r bachgen ddoe

Gave-**3sg** the teacher book to-the boy yesterday

The teacher gave a book to the boy yesterday

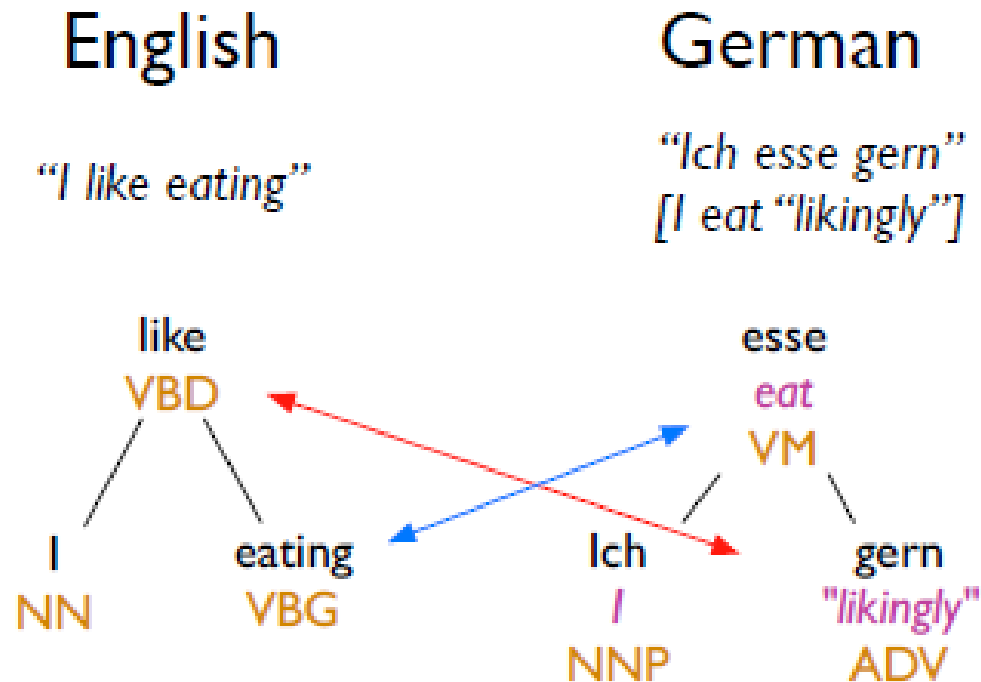


Enriching IGT: parse the translation and project the parse tree



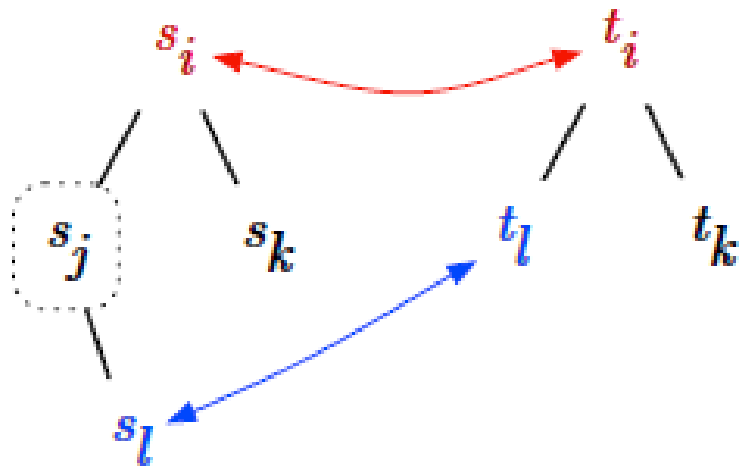
Translation divergence

- Dorr (1994) outlines seven types of divergence that may occur between languages.

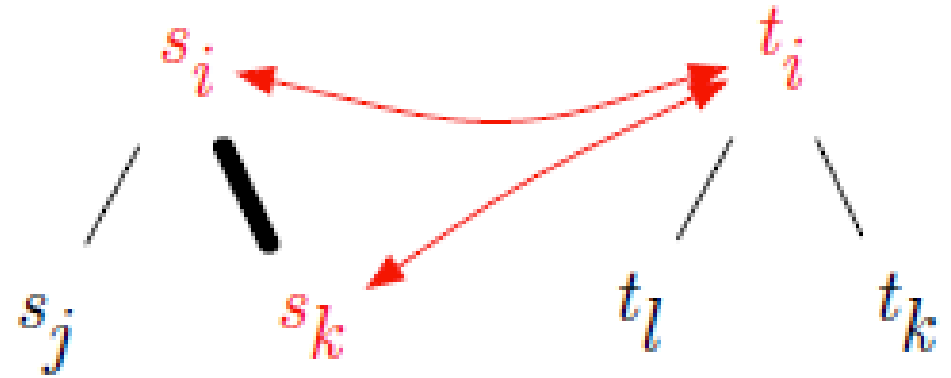


→ What are common divergence types and can they be detected automatically?

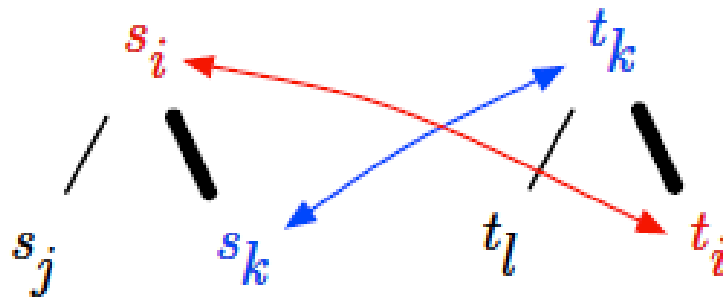
Alignments that cause divergence



(1) Unaligned words

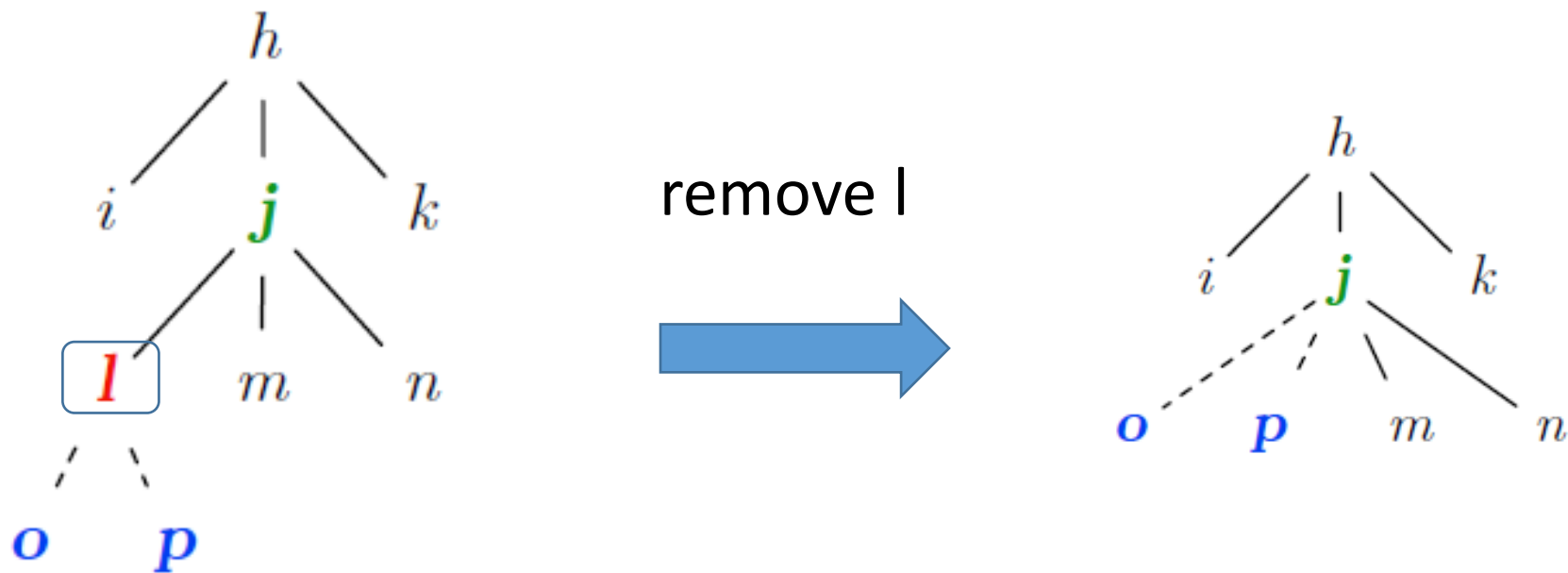


(2) many-to-one alignment

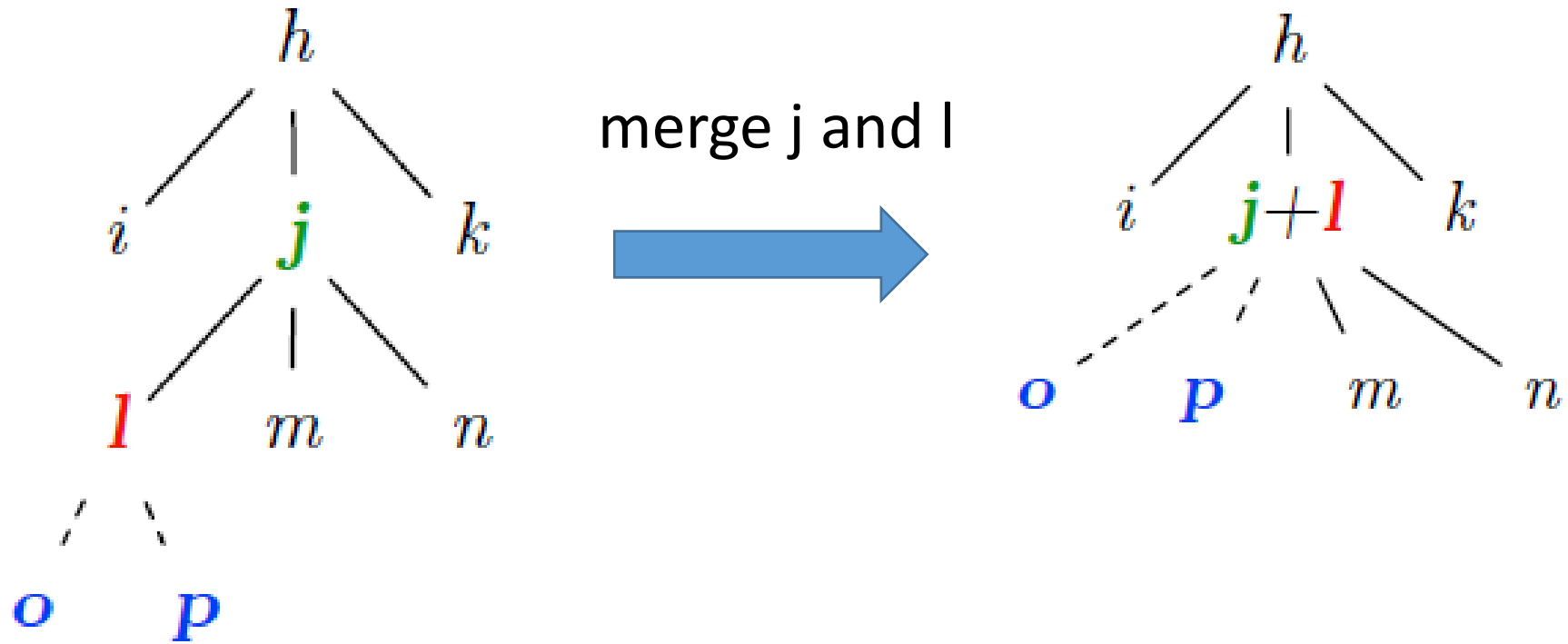


(3) head switching

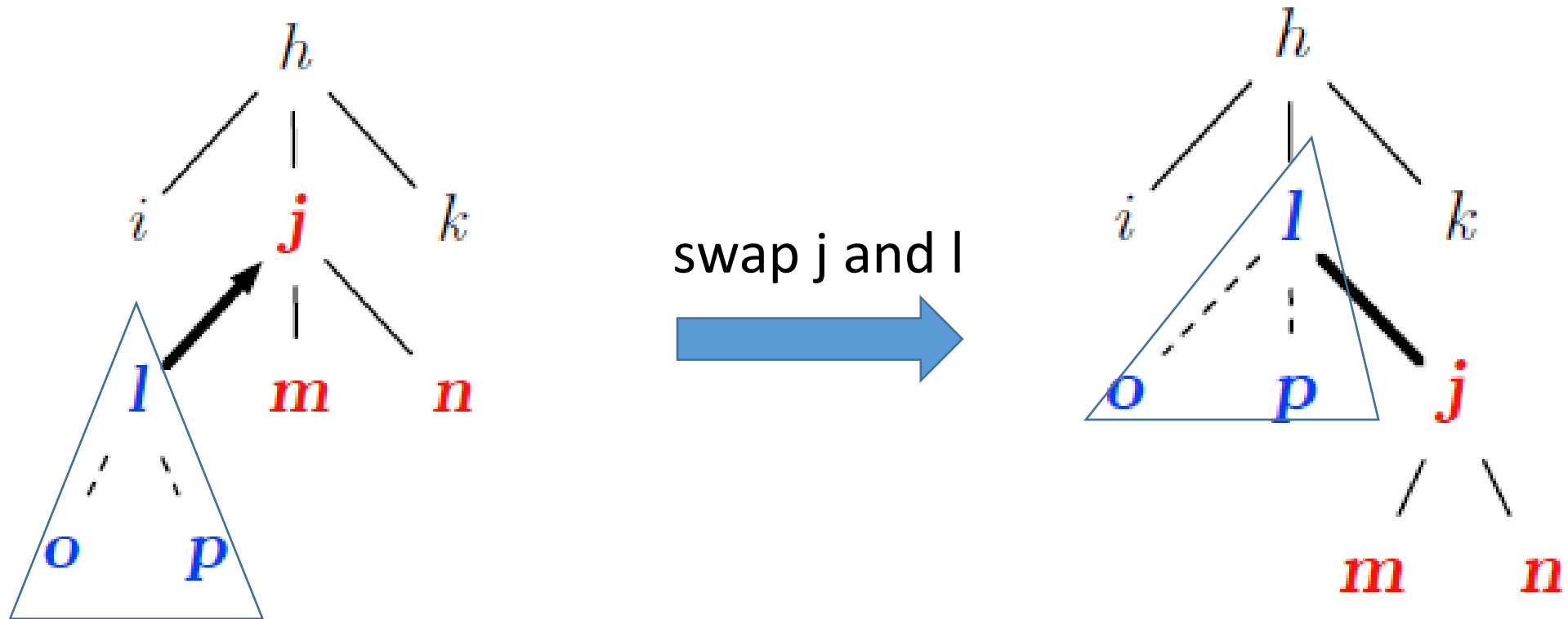
“Remove” operation for unaligned words



“Merge” operation for many-to-one alignment



“Swap” operation for head switching



Transforming a tree pair with three operations

- Input:
 - tree pair (S, T)
 - word alignment between words in S and T
- Output: transformed tree pair (S', T')
- Steps:
 - For unaligned words in S and T, apply "Remove"
 - For many-to-one alignment, apply "Merge"
 - For head switching, apply "Swap"

Unaligned words

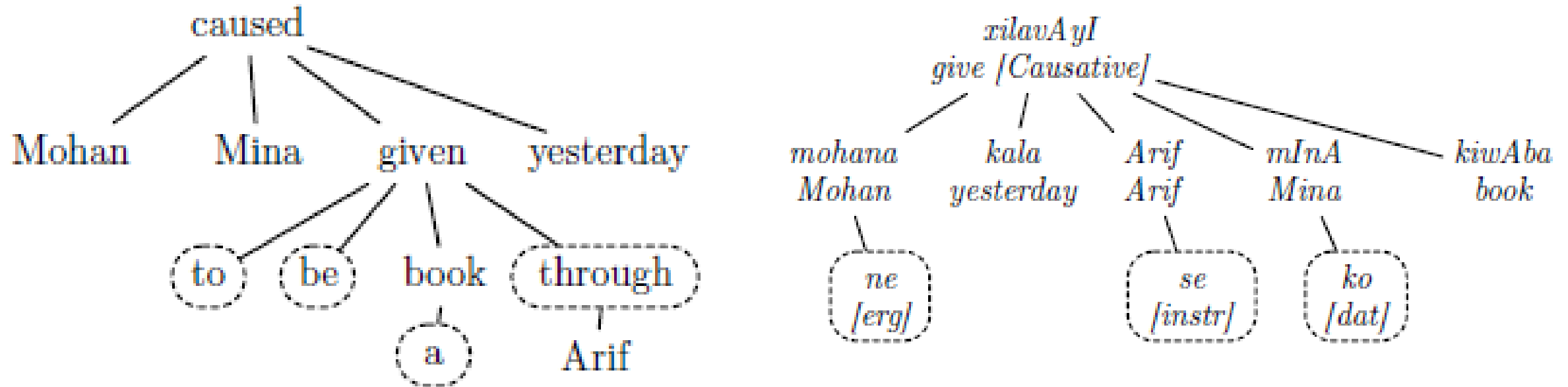
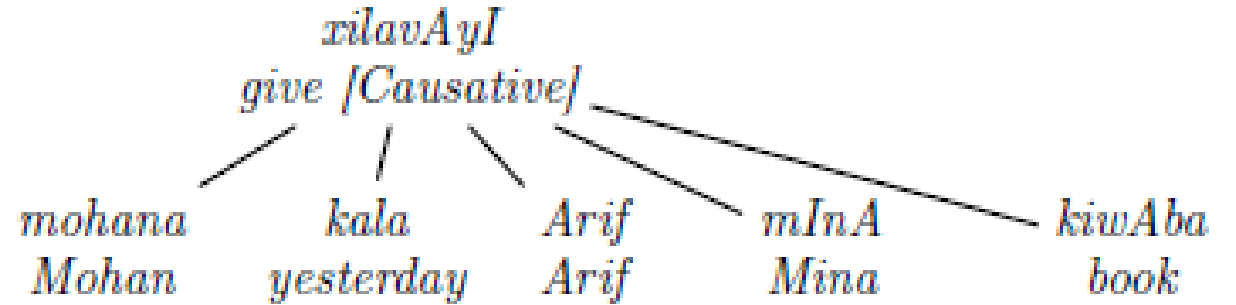
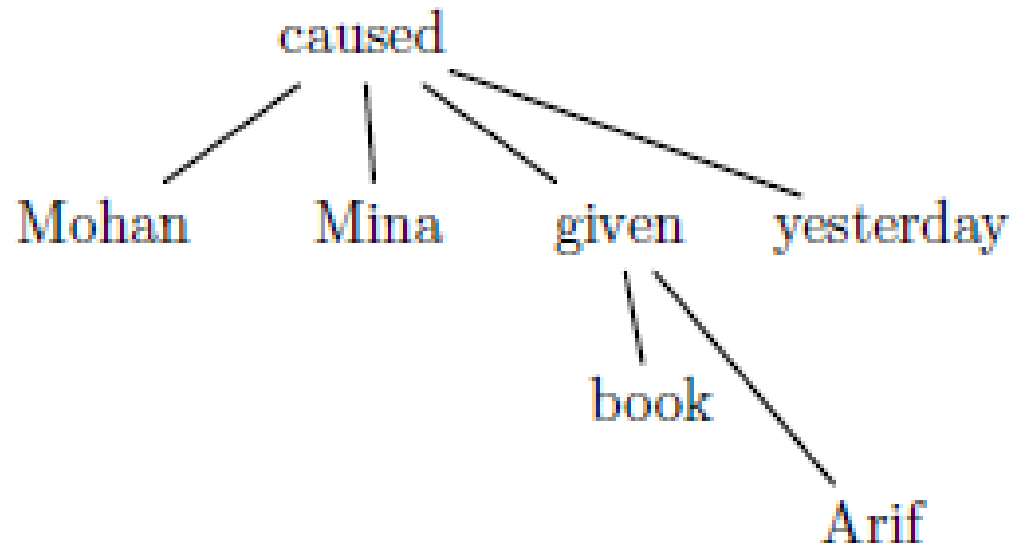
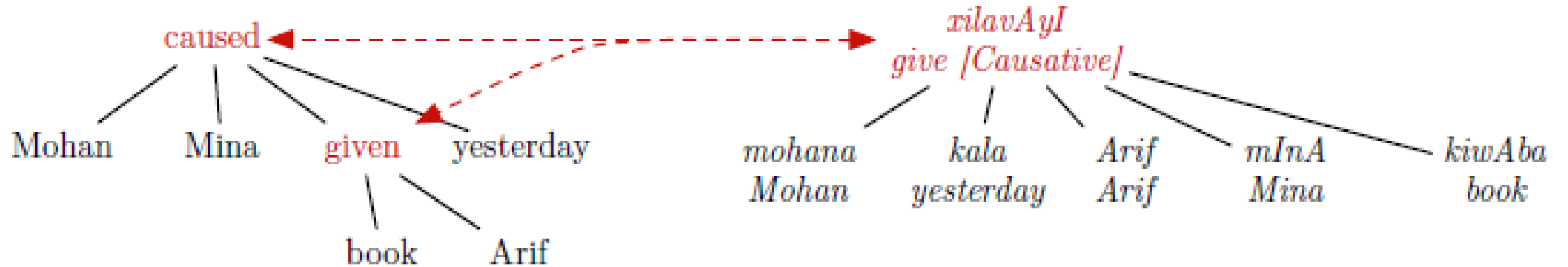


Figure: The trees for “Mohan caused Mina to be given a book through Arif yesterday” and its Hindi counterpart

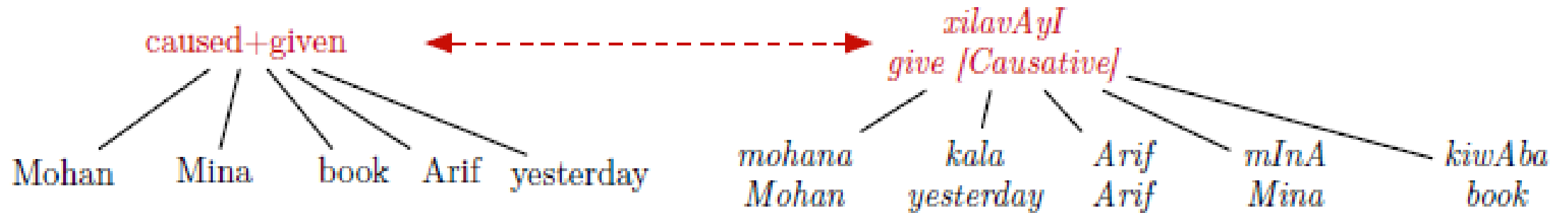
After removing unaligned words



Many-to-one alignment



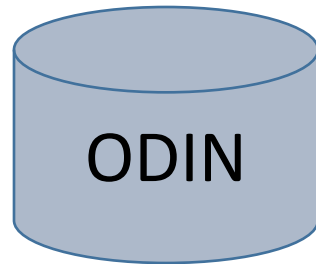
After applying the “merge” operation



Improving syntactic projection (Georgi et al., 2014)

- Syntactic projection is error-prone due to language divergence
- To study divergence:
 - We define a metric to compare structural similarity between a treebank pair
 - We identify three common divergence types and define a tree operation for each type
 - Experiments demonstrate the effect of these operations on matching percentage
- Improve projection accuracy:
 - Learn divergence patterns from a small training corpus (a parallel treebank which can come from enriched IGT)
 - Apply the patterns to the output of the basic projection algorithm
- The whole process is automatic and does not require language-specific knowledge.

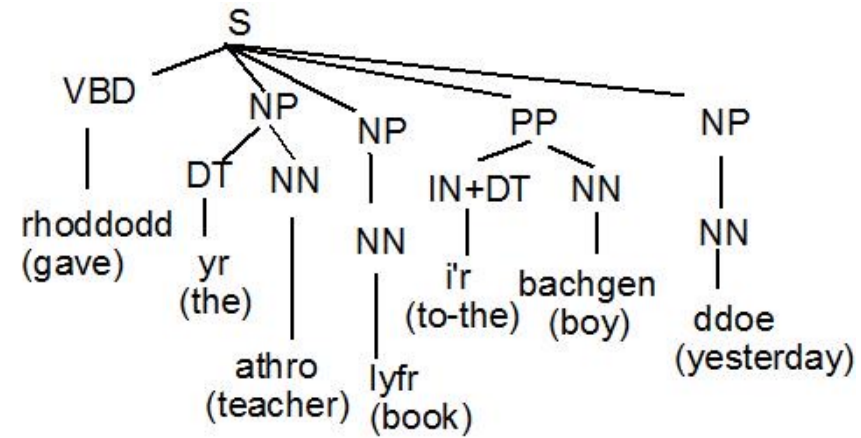
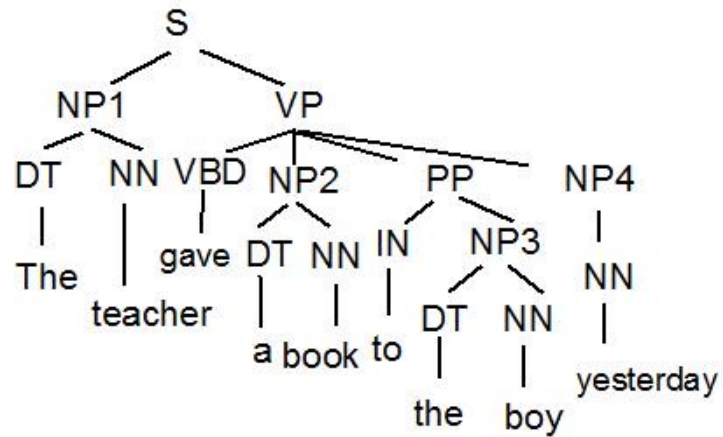
Automatically answer linguistic questions (Lewis and Xia, 2008)



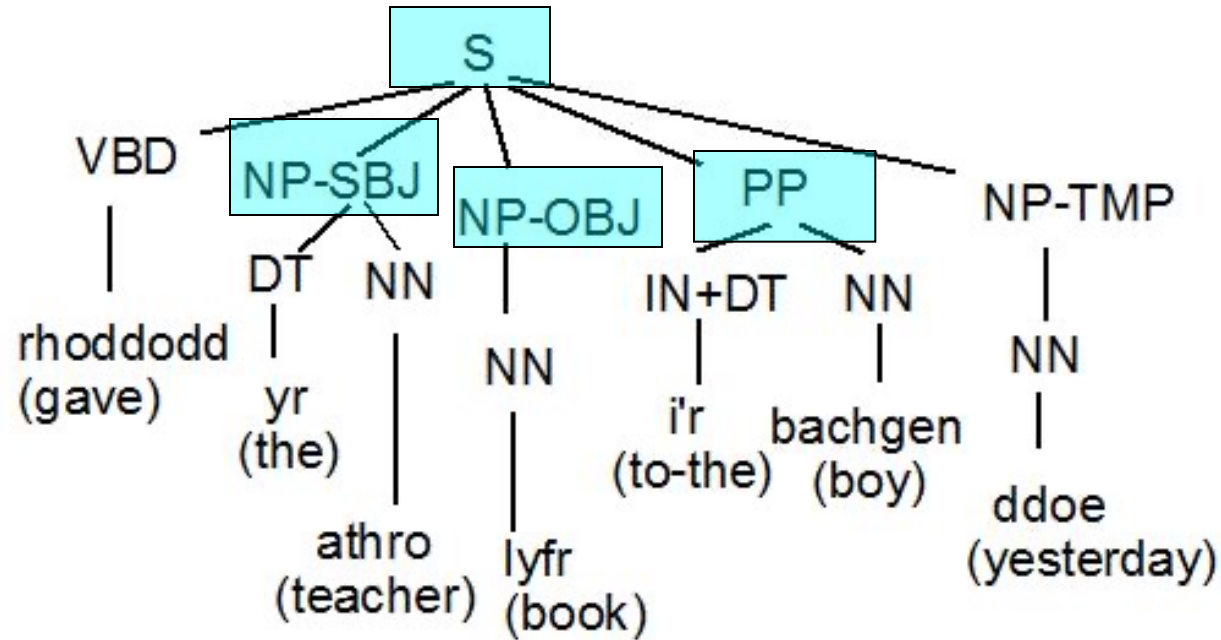
	Q ₁	Q ₂	...
L ₁			
L ₂			
...			

WALS (World Atlas of Language Structures)

Enriching IGT: parse the translation and project the parse tree



Extracting context-free rules



$S \rightarrow VBD \ NP\text{-}SBJ \ NP\text{-}OBJ \ PP \ NP\text{-}TMP$

$NP\text{-}SBJ \rightarrow DT \ NN$

$NP\text{-}OBJ \rightarrow NN$

$PP \rightarrow IN+DT \ NN$

Extracted context-free grammar

- $S \rightarrow V \text{ NP-SBJ NP-OBJ}$ 0.40
- $S \rightarrow V \text{ NP-SBJ}$ 0.30
- $S \rightarrow V \text{ NP-SBJ NP-OBJ PP}$ 0.10
- ...

- $\text{NP} \rightarrow \text{DT NN}$ 0.51
- $\text{NP} \rightarrow \text{NN}$ 0.26
- $\text{NP} \rightarrow \text{ADJ NN}$ 0.13
- ...

➔ The language seems to be VSO, and the order between DT and NN is DT-NN.

Answering questions in language profile

- 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 <i>a, the</i>
358	Demonstrative, Noun	Order of Nouns and Demonstrative Determiners (<i>this, that</i>)
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

Experiment #1

10 languages
13 questions

Parameter	Accuracy
WOrder	90%
VP-OBJ	60%
DT-NN	80%
Dem-NN	90%
JJ-NN	100%
PRP\$-NN	80%
Poss-NN	70%
P-NP	90%
number	70%
case	80%
T/A	80%
Def	100%
Indef	90%
Average	83%

Experiment #2

- Project Structures for 97 languages
- Determine value of the word order parameter for each language (e.g., SVO, SOV, etc.)
- How much data is required for accurate answers?

Results

- Accuracy: For 69 of the 97 languages, WOrder was accurately determined
- Confusion matrix:

System output

		System output			
		SVO	SOV	VSO	VOS
Truth	SVO	32	8	0	9
	SOV	2	33	0	6
	VSO	2	2	3	4
	VOS	0	0	0	1

Results for Experiment #2

- Accuracy improved as # of IGT instances increased

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

Error analysis

- Insufficient Data
 - Ex: most VSO languages had less than 10 instances.
- Skewed or Inaccurate Data: (a.k.a. **IGT bias**)
 - Ex: Cantonese is SVO and had over 73 instances in ODIN, but one source had a large number of skewed SOV instances.
- Projection errors: (a.k.a. **English bias**)
 - A combination of errors in English parse tree, word alignment, and projection
- Free Constituent Order:
 - Free word order is more difficult to assign a value to than a fixed word order.
 - Even with “Fixed” Word Order languages, word order can be flexible (and degree can be flexible cross-linguistically)

What the results show

- We can fairly accurately discern values for several typological parameters.
- Larger samples overcome the effects of the IGT bias and the English bias.
- We can do this across many languages *automatically*.

Outline

- Natural Language Processing (NLP) and linguistics
- The RiPLes project
 - Motivation
 - ODIN: Collecting language data from the Web
 - INTENT: Enrich the data
- **The AGGREGATION project**

The AGGREGATION project

- The LinGO Grammar Matrix (Bender et al., 2002) is cross-linguistic grammar resource designed to facilitate the creation of machine-readable, linguistically motivated grammars for any human language.
 - A core grammar and a series of libraries
 - A customization system
- Goals of the AGGREGATION project:
 - To generate grammar (fragments) automatically using information from IGT and other types of language data
 - To bring the benefits of grammar engineering to descriptive and documentary linguists.

Grammar Matrix customization page

<http://matrix.ling.washington.edu/customize/matrix.cgi>

Word order:

- SOV
- SVO
- VSO
- OSV
- OVS
- VOS
- V-final
- V-initial
- free (pragmatically determined word order)
- finite verb or auxiliary in second position, else free word order
- finite verb second, non-finite verb clause-finally

Enriched IGT

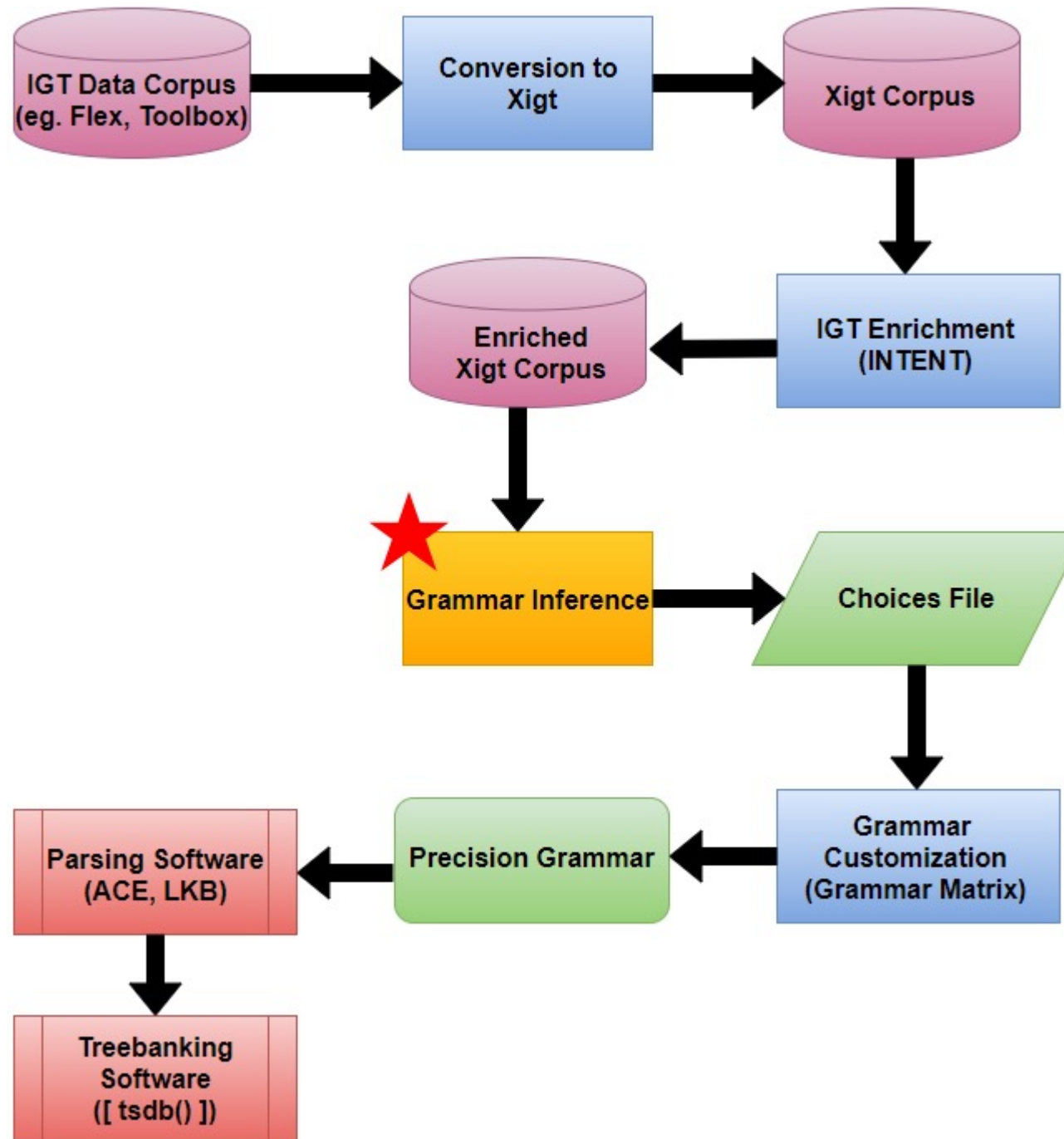
- Original IGT: Language line, gloss line, translation line
- Enriched IGT:
 - words and morphemes in the language line
 - Words and morphemes in the gloss line
 - Words in the translation line
 - Alignment between words (and morphemes) in the language line and gloss line
 - Alignment between the words in the gloss line and the translation line

 - Part-of-speech (POS) tags for the words in the translation line
 - Phrase structure (PS) and dependency structure (DS) for the translation line
 - Projected POS tags for the words in the language line
 - Projected PS and DS for the language line

XIGT: an eXtensible representation for IGT (Goodman et al, 2015)

- Properties of IGT:
 - Stand-off: the ability to deploy an annotation without changing the original data
 - Incrementality: allowing for incremental development of analyses
 - Extensibility: easy to add additional annotation “tiers”
 - Complex alignments: allow complex alignments between annotation tiers
 - ID-reference annotation:
 - ...
- For more info and source code, see <https://github.com/xigt/xigt>

AGGREGATION:



Case studies

- Inferring case systems (Howell et al., 2017)
- Computational support for finding word classes in Abui (Zamaraea et al., 2017)
- Study lexical classes in Chintang (Zamaraea et al., 2019)
- Extracting typological and lexical properties (Howell, 2020)

Conclusion

- NLP and linguistics are closely related, and we are interested in exploring ways that the two fields can benefit each other.
- RiPLes is a project that uses NLP techniques to
 - collect IGT data from the Web → ODIN database
 - project information from resource-rich languages to resource-poor languages → enriched IGT
 - create language profiles (e.g., grammar fragments) from the enriched IGT
 - Use enriched IGT to bootstrap NLP tools (e.g., parsers) to process more language data

Conclusion (cont)

- The AGGREGATION project:
 - Built on top of RiPLes and LINGO Grammar Matrix projects
 - XIGT was proposed as a data format to store enriched IGT.
 - The main component is grammar inference (e.g., word order, case system, lexical classes).
 - The output of the system can help descriptive and documentary linguists.
- NLP can help linguistic studies:
 - NLP can be used to collect and enrich language data
 - Enriched data can be used to infer linguistic information or to bootstrap NLP systems so that more data can be processed automatically.

More information

- ODIN v2.1 download: https://uakari2.ling.washington.edu/corpus/odin/v2_1/09384dc6/
- ODIN and XIGT source code: <https://github.com/xigt>
- A demo for an IGT Editor: <http://editor.xigt.org/user/demo>
- AGGREGATION: <http://depts.washington.edu/uwcl/aggregation/>

References: the RiPLes project

- Ryan Georgi, 2016. From Aari to Zulu: Massively Multilingual Creation of Language Tools Using Interlinear Glossed Text. PhD thesis, University of Washington.
- Ryan Georgi, Michael Wayne Goodman, and Fei Xia, 2016. "A Web-framework for ODIN Annotation", in Proceedings of ACL-2016 System Demonstrations, pp 31-36, Aug 7-10, Berlin, Germany.
- Fei Xia, William D. Lewis, Michael W. Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily Bender, 2016. "Enriching a Massively Multilingual Database of Interlinear Glossed Text", Journal of Language Resources and Evaluation (LRE), 50(2): 321-349.
- Ryan Georgi, Fei Xia, and William D. Lewis, 2014. Capturing Divergence in Dependency Trees to Improve Syntactic Projection, Journal of Language Resources and Evaluation (LRE).
- William Lewis and Fei Xia, 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages, Journal of Literary and Linguistic Computing (LLC), 25(3):303-319.
- Fei Xia, Carrie Lewis and William D. Lewis, 2010. The Problems of Language Identification within Hugely Multilingual Data Sets, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), pages 2790-2797, Valletta, Malta, May 19-21, 2010.
- Fei Xia, William Lewis and Hoifung Poon, 2009. Language ID in the Context of Harvesting Language Data off the Web, Proceedings of the 12th Conference of the European Chapter of the ACL (EACL-2009), pages 870-878, Athens, Greece, March 30 - April 3, 2009.
- William Lewis and Fei Xia, 2008. Automatically Identifying Computationally Relevant Typological Features, Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-2008), pages 685-690, Hyderabad, India, Jan 7-12, 2008.

References: the AGGREGATION project

- Howell, Kristen. 2020. Inferring Grammars from Interlinear Glossed Text: Extracting Typological and Lexical Properties for the Automatic Generation of HPSG Grammars. PhD thesis, University of Washington.
- Zamaraeva, Olga, Kristen Howell and Emily M. Bender. 2019. Handling Cross-cutting Properties in Automatic Inference of Lexical Classes: A Case Study of Chintang. *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu, HI. Honolulu, HI. pp.28-38.
- Zamaraeva, Olga, Kristen Howell and Emily M. Bender. 2019. Modeling Clausal Complementation for a Grammar Engineering Resource. *Proceedings of the Society for Computation in Linguistics* Vol. 2, Article 6.
- Zamaraeva, Olga, František Kratochvíl, Emily M. Bender, Fei Xia and Kristen Howell. 2017. Computational Support for Finding Word Classes: A Case Study of Abui. In *Proceedings of ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages*, ICLDC 2017, Honolulu Hawai`i.
- Howell, Kristen, Emily M. Bender, Michael Lockwood, Fei Xia and Olga Zamaraeva. 2017. Inferring Case Systems from IGT: Impacts and Detection of Variable Glossing Practices. In *Proceedings of ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages*, ICLDC 2017, Honolulu Hawai`i.
- Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender, 2015. Xigt: Extensible Interlinear Glossed Text for Natural Language Processing, in *Language Resources and Evaluation*, 49(2):455-485.