Extracting granular information on habitats and reproductive conditions of Dipterocarps through pattern-based literature analysis

Roselyn Gabud[1, 2] *and Riza Batista-Navarro*[2,3]

[1]University of the Philippines Diliman, Philippines; [2]University of the Philippines Los Baños, Philippines; [3]The University of Manchester, UK

Roselyn S. Gabud

rsgabud@up.edu.ph

Department of Computer Science

University of the Philippines Diliman, Los Baños

27 September 2018

ICEI2018 Jena, Germany

What are Dipterocarps?

- *Dipterocarpaceae*
- medium to large forest trees, skeletal backbone of lowland tropical forests
- ~65 species in 6 genera in the Philippines, more than 65% are endemic
- economically and ecologically important, e.g., timber value
  - ➔ exploited and affected by decline in forest cover:

Challenge: Reproduction of Dipterocarps

1. Long-term (temporal)
2. Broad-scale (geographical)

Photo by: Edwino S. Fernando. 07 December 2006.

Aims and Objectives

- **Aim**: To develop literature mining methods to automatically extract information relevant to the distribution and reproductive cycle of dipterocarps
  - in order to help predict the likelihood of their regeneration, and
  - subsequently make informed decisions regarding species for reforestation.

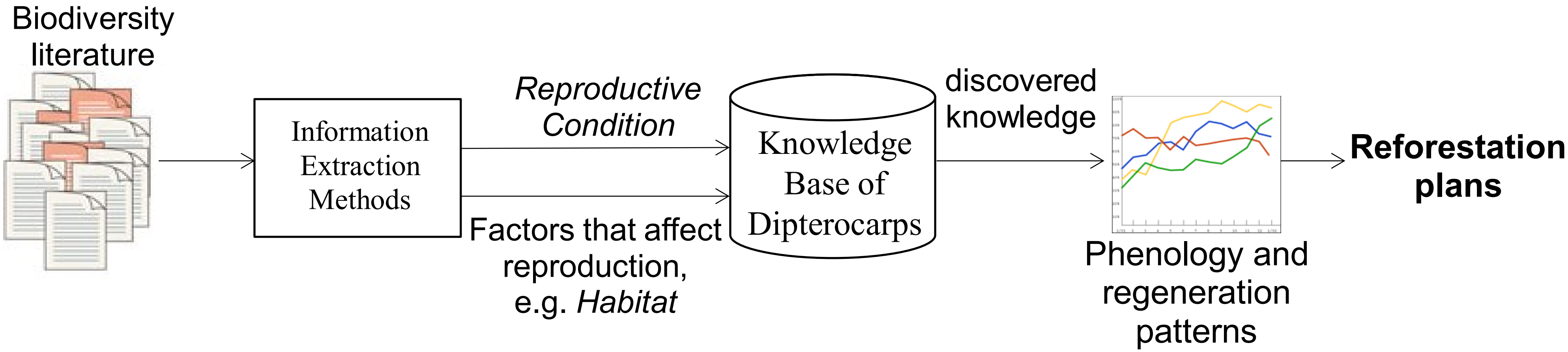

Biodiversity literature
Information Extraction Methods
Reproductive Condition
Factors that affect reproduction, e.g. *Habitat*
Knowledge Base of Dipterocarps
discovered knowledge
Phenology and regeneration patterns
Reforestation plans

Figure 1. An overview of the research aims and objectives.

DipteroMine Corpus

Journals

155 abstract length documents from:

- Journal of Tropical Ecology
- Journal of Ecology
- Journal of Biosciences
- Forest Ecology and Management

	Concept	Description	Example
1	Habitat	Environments in which organisms live.	In the [lowland mixed dipterocarp forests] of Borneo the Dipterocarpaceae can comprise roughly 107 of species …
2	Geographical Location	Any identifiable point or area in the planet. (countries, major bodies of water, named landforms, etc).	The main observation site was conserved forest at [Dongmakhai] ([18deg20 ′ 03 ″ N , 102deg30 ′ 5 ″ E] , 190 m a.s.l.)
3	Reproductive Condition	Indicators of the specimens' reproductive condition.	There were two [flowerings] in March to May , and one in August during this period .
4	Temporal Expression	Spans of text pertaining to points in time.	Most fruit fall occurred from the [end of July] to [mid-August].

Gabud, R.S., et al. Understanding mass flowering of dipterocarps through semantic occurrence information extraction. TDWG 2016 Annual Conference.

DipteroMine Corpus

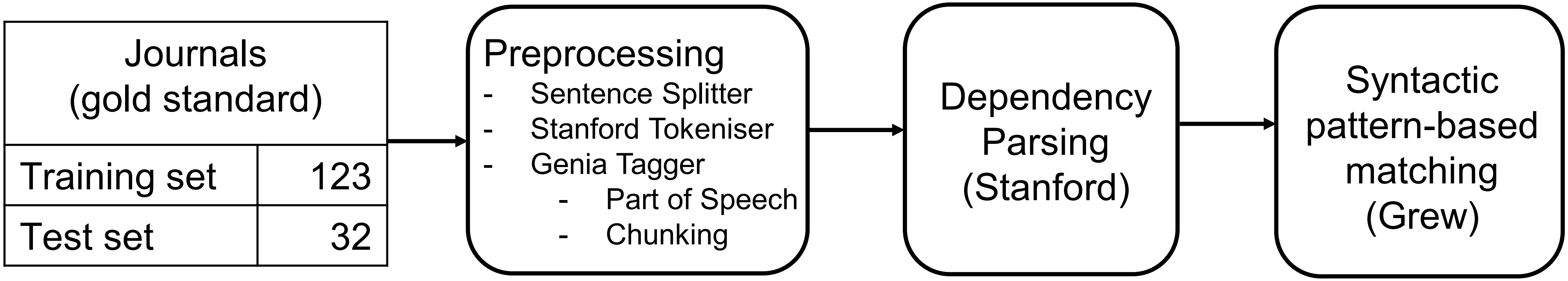
Double annotation: 79
Single annotation: 76

Inter-Annotator Agreement

	size	F score
Geographical Location	711	92%
Habitat	475	75%
Temporal Expression	787	91%
Reproductive Condition	539	64%

Gabud, R.S., et al. Understanding mass flowering of dipterocarps through semantic occurrence information extraction. TDWG 2016 Annual Conference.

Methodology



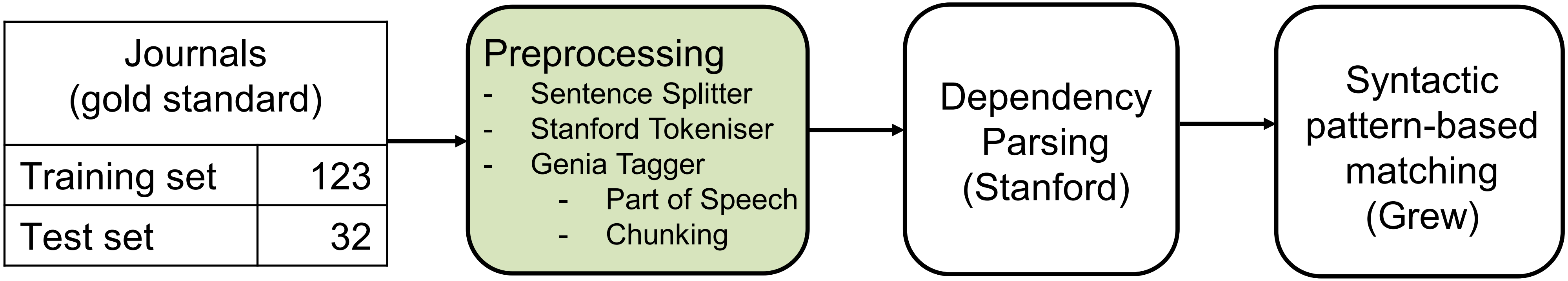
Journals (gold standard)	
Training set	123
Test set	32

Preprocessing
- Sentence Splitter
- Stanford Tokeniser
- Genia Tagger
    - Part of Speech
    - Chunking

Dependency Parsing (Stanford)

Syntactic pattern-based matching (Grew)

Methodology



Journals (gold standard)	
Training set	123
Test set	32

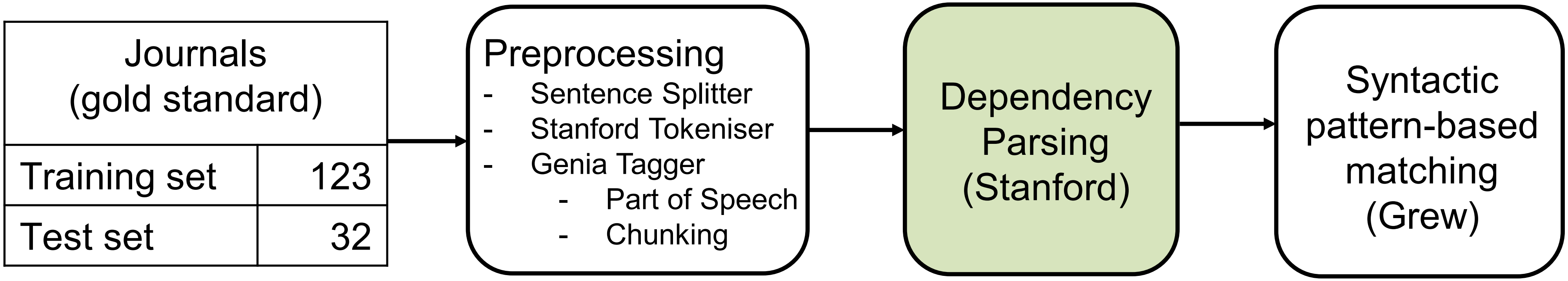
Preprocessing

- Sentence Splitter
- Stanford Tokeniser
- Genia Tagger
  - Part of Speech
  - Chunking

Dependency Parsing (Stanford)

Syntactic pattern-based matching (Grew)

Methodology



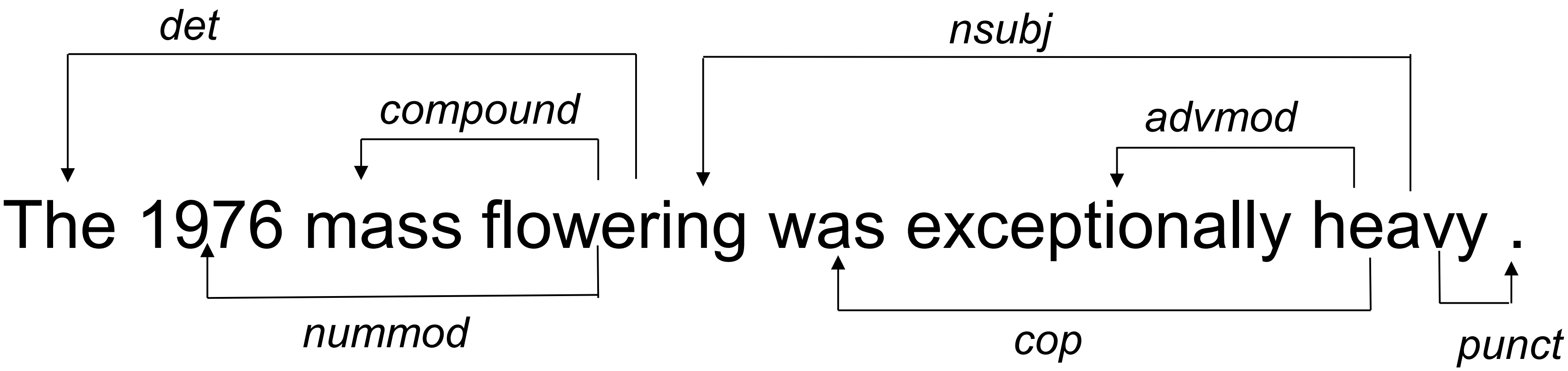
Journals (gold standard)	
Training set	123
Test set	32

Preprocessing
- Sentence Splitter
- Stanford Tokeniser
- Genia Tagger
    - Part of Speech
    - Chunking

Dependency Parsing (Stanford)

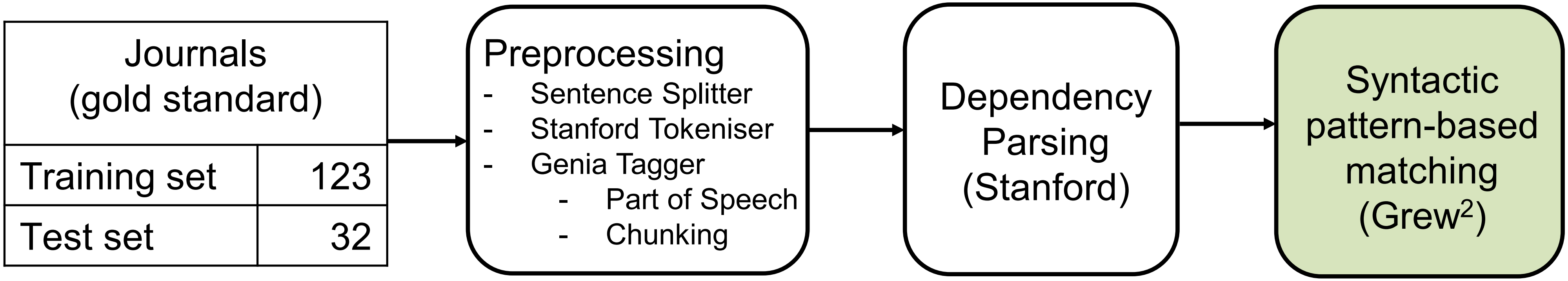
Syntactic pattern-based matching (Grew)

Stanford dependencies[1] provides a representation of grammatical relations between words in a sentence.



[1] https://nlp.stanford.edu/software/stanford-dependencies.shtml

Methodology



Journals (gold standard)	
Training set	123
Test set	32

Preprocessing
- Sentence Splitter
- Stanford Tokeniser
- Genia Tagger
  - Part of Speech
  - Chunking

Dependency Parsing (Stanford)

Syntactic pattern-based matching (Grew[2])

Grew

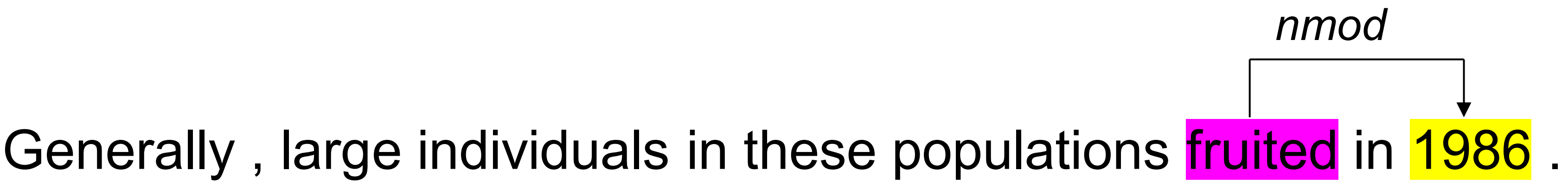
- is a Graph Rewriting tool dedicated to applications in Natural Language Processing (NLP).
- lets the user search for a given pattern in a corpus of syntactic structures.

[2] http://grew.fr/

Syntactic Pattern-Based Matching

1. Direct relationship between entities.

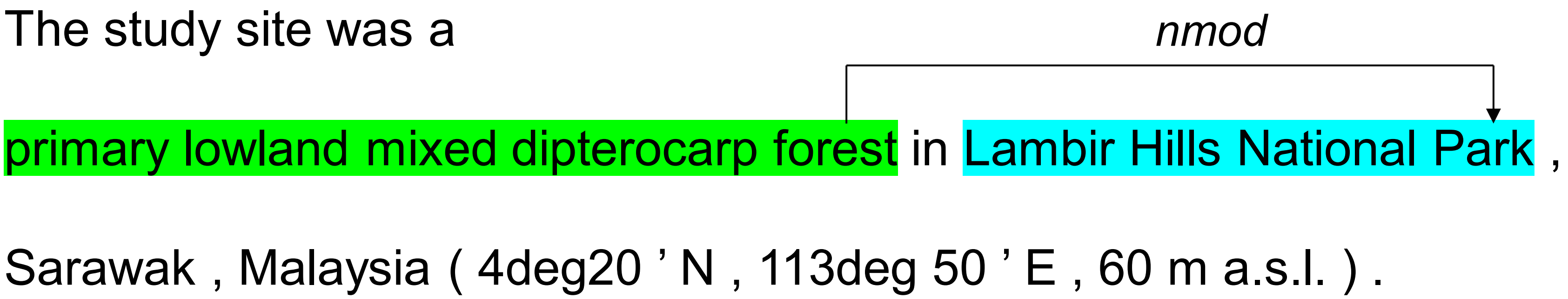
Reproductive Condition → Temporal Expression



Generally , large individuals in these populations fruited in 1986 .

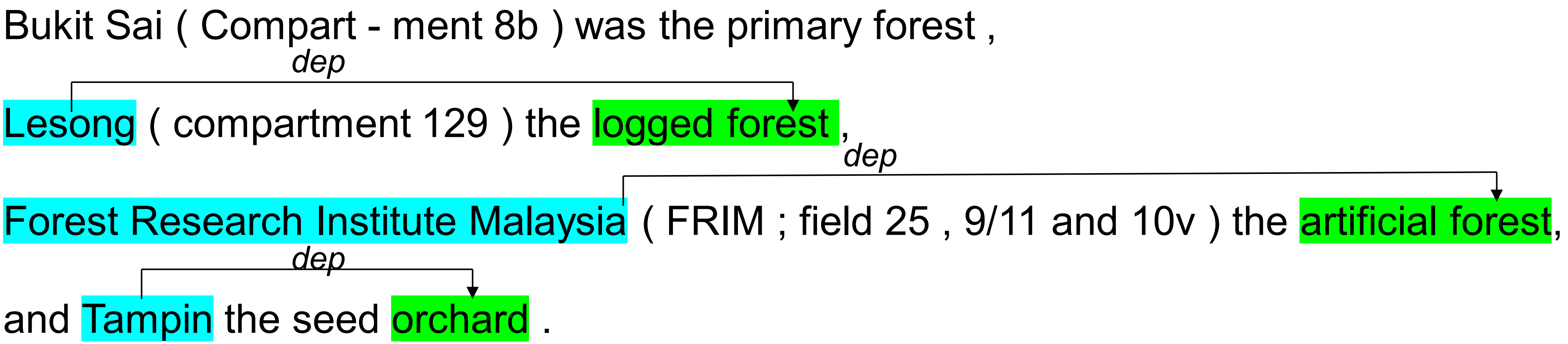
Syntactic Pattern-Based Matching

Habitat → Geographical Location



The study site was a primary lowland mixed dipterocarp forest in Lambir Hills National Park , Sarawak , Malaysia (4deg20 ’ N , 113deg 50 ’ E , 60 m a.s.l.) .

Geographical Location → Habitat

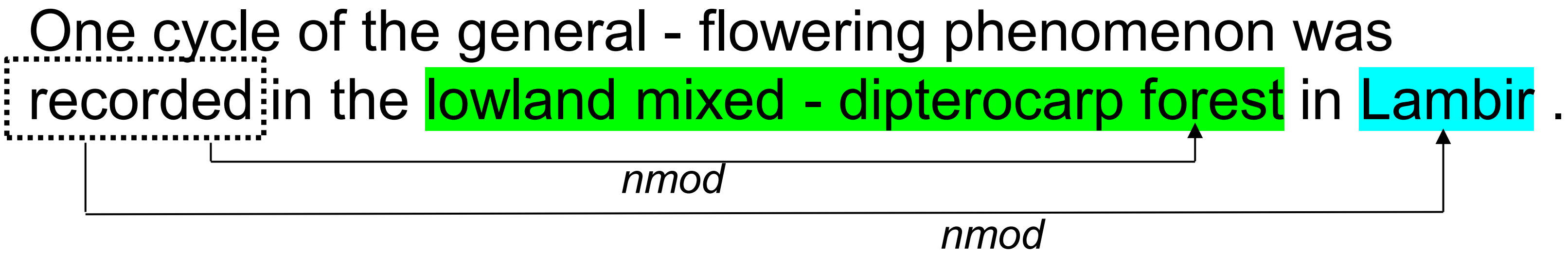


Bukit Sai (Compart - ment 8b) was the primary forest , Lesong (compartment 129) the logged forest , Forest Research Institute Malaysia (FRIM ; field 25 , 9/11 and 10v) the artificial forest, and Tampin the seed orchard .

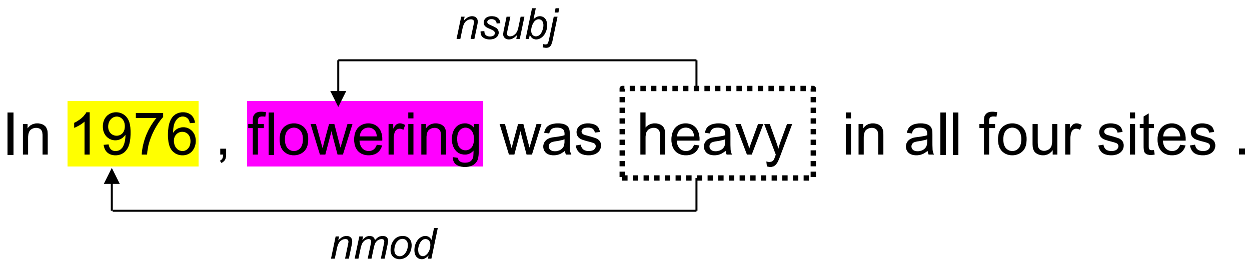
Syntactic Pattern-Based Matching

2. Entities have a common root.

N→ Habitat
N→ Geographical Location


One cycle of the general - flowering phenomenon was recorded in the lowland mixed - dipterocarp forest in Lambir .

N→ Reproductive Condition
N→ Temporal Expression

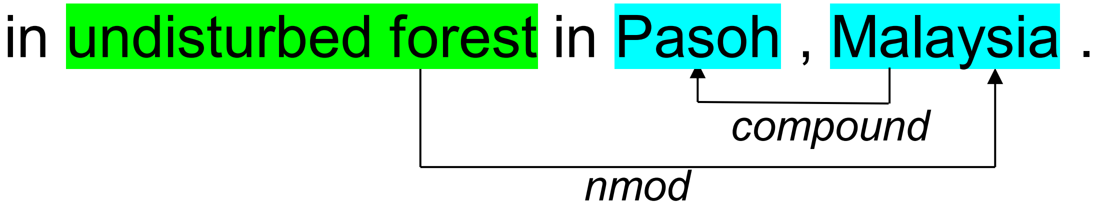

In 1976 , flowering was heavy in all four sites .

Syntactic Pattern-Based Matching

3. Entities are linked by 1 or more tokens (words).

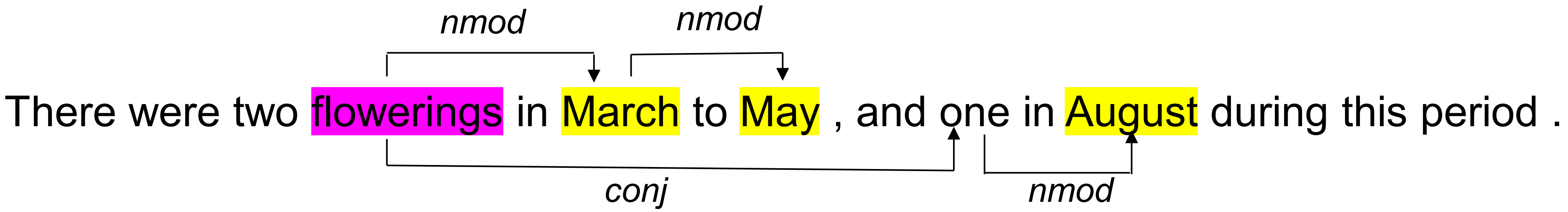
Habitat → N → Geographical Location

Appanah and Rasol (1990) reported that mean dbh of fruiting dipterocarp trees was 70.2 cm



in undisturbed forest in Pasoh , Malaysia .

Reproductive Condition → N→ Temporal Expression



There were two flowerings in March to May , and one in August during this period .

Sample relations extracted

Habitat	Geographical Location
lowland dipterocarp forest	Sarawak
swamps	northwest Borneo
tropical forests	southeast Asia
lowland dipterocarp forest	Lambir Hills National Park
logged forest	Lesong
fresh water swamps	Sabah

Reproductive Condition	Temporal Expression
flowering	end of November 2001
flowering	end of August 2001
mast fruiting	Aug-96
mass flowering	1976
mass flowering	1955
flowered	Jul-66

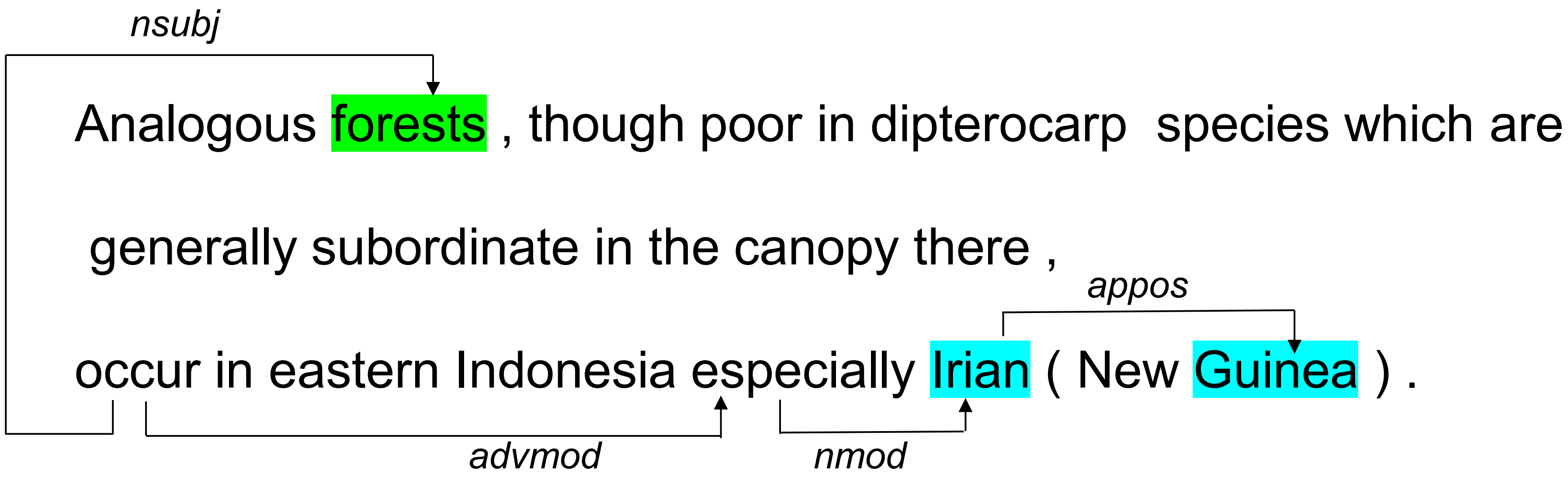
Evaluation

Relation Type	Method	Relevant relations	TP	FP
Habitat – Geographical Location	Co-occurrence	47	47	26
	Relation extraction	47	38	2
Reproductive Condition – Temporal Expression	Co-occurrence	139	139	144
	Relation extraction	139	90	11

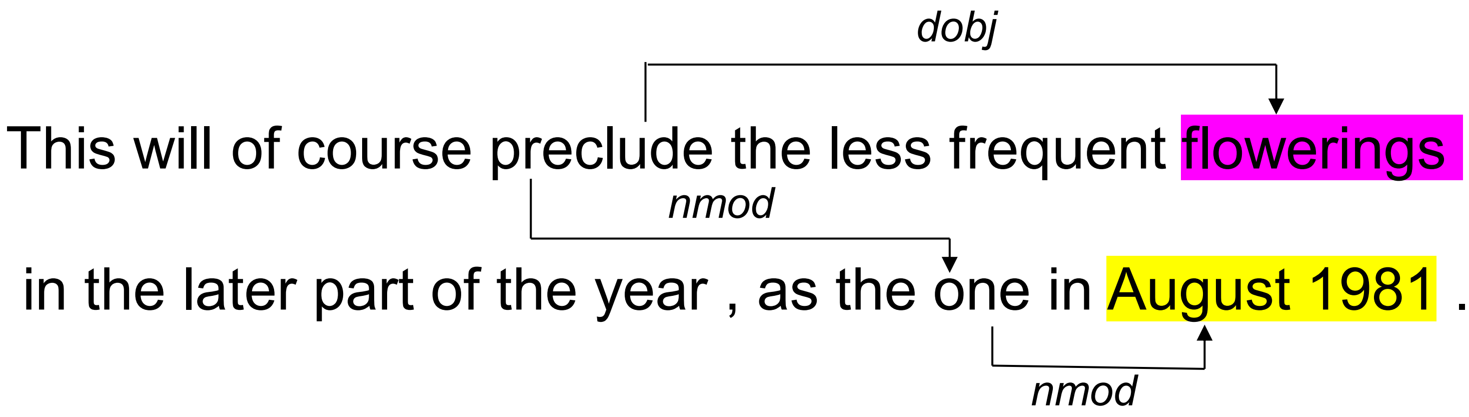
Relation Type	Method	Precision	Recall	F-score
Habitat – Geographical Location	Co-occurrence	64.38%	100.00%	78.33%
	Relation extraction	95.00%	80.85%	87.36%
Reproductive Condition – Temporal Expression	Co-occurrence	49.12%	100.00%	65.88%
	Relation extraction	89.11%	64.75%	75.00%

$$F = \frac{2 * pre * recall}{pre + recall}$$

Examples of missed relations



Analogous forests , though poor in dipterocarp species which are generally subordinate in the canopy there , occur in eastern Indonesia especially Irian (New Guinea) .



This will of course preclude the less frequent flowerings in the later part of the year , as the one in August 1981 .

Ongoing Work

- Consider the presence of modifiers between a common root of entities.
- Curate a database of dipterocarp occurrences using relation extraction based on syntactic pattern matching, i.e. integration of text-mined information (e.g., Habitat – Geographical Location and Reproductive Condition – Temporal Expression relationships) with primary data (e.g., occurrence data from GBIF).

Acknowledgements

PhD Supervisors

Riza Batista-Navarro, PhD
National Center for Text Mining
University of Manchester
Institute of Computer Science
University of the Philippines Los Baños

Vladimir Mariano, PhD
Centre of Technology
RMIT University, Vietnam

Eduardo Mendoza, PhD
Max Planck Institute of Biochemistry
Martinsried, Germany

Portia Lapitan, PhD
Nelson Pampolina, PhD
College of Forestry and Natural Resource
University of the Philippines Los Baños

Corpus funding source: COPIOUS
British Council under the Newton Fund
Prof. Sophia Ananiadou
National Center for Text Mining
University of Manchester

Collaborators:

Nhung Nguyen, PhD
National Center for Text Mining
University of Manchester

Sandra Yap, PhD
Fareastern University Manila

Research Assistants:
John Marc Cho Santos

Thank you!

Questions?

Roselyn S. Gabud
rsgabud@up.edu.ph
Department of Computer Science
University of the Philippines Diliman, Los Baños