Corpora in NLP

Wirote Aroonmanakun

Corpus use

Lexicography

Linguistics research

Language teaching

Translation studies

Natural language processing

Good corpus

Representative

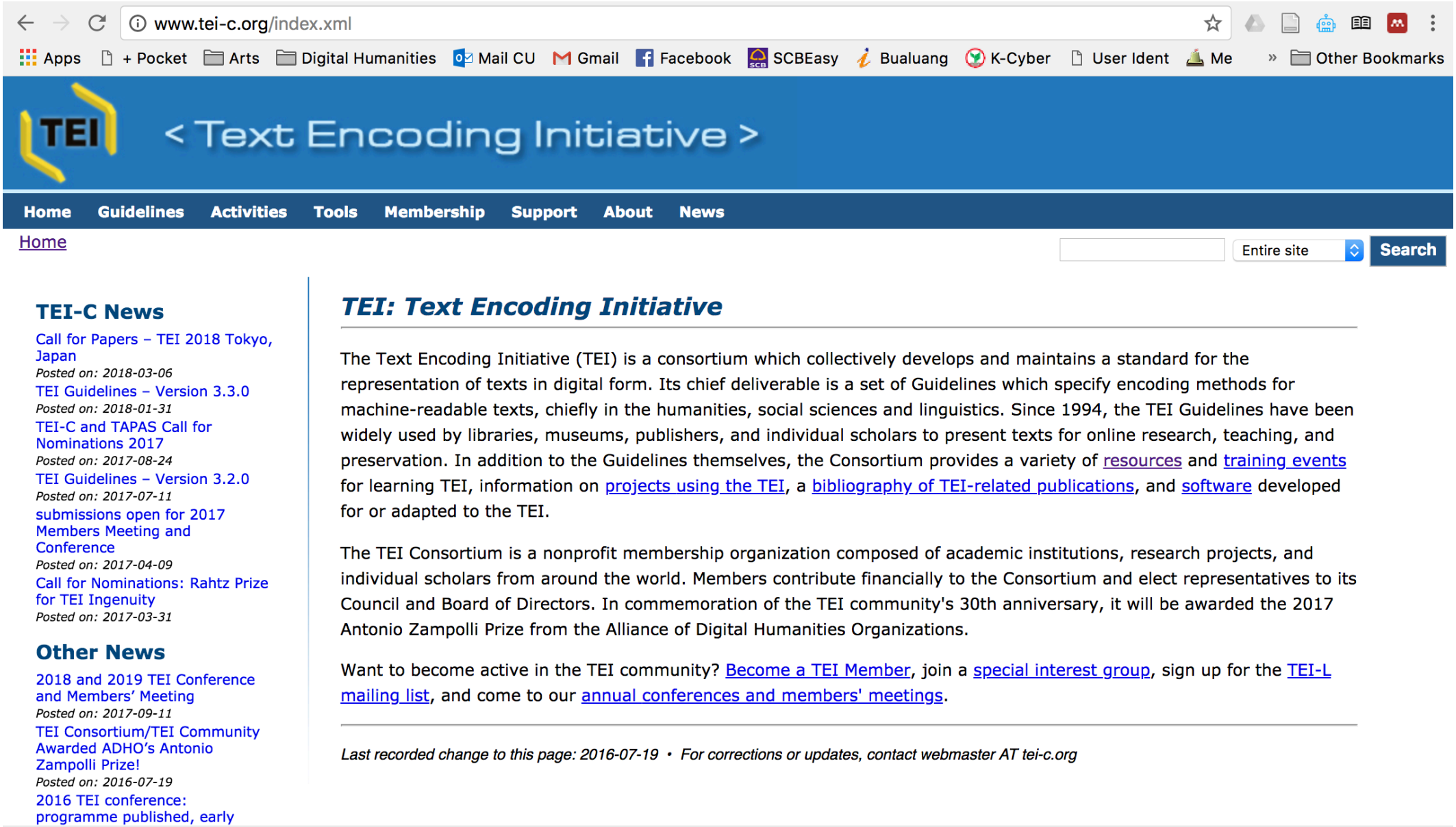
Balance

Size

Annotation

Developing Linguistic Corpora: a Guide to Good Practice (http://ota.ox.ac.uk/documents/creating/dlc/)

Text Encoding Initative



Front Matter

Back Matter

Text Body

TEI sourcecode

- Getting and Using the TEI Sources.
- TEI GitHub Repository
- Bug Reports, Feature Requests, etc.

http://www.tei-c.org/index.xml

CoNLL-U Format (CoNLL shared task)

```
# sent id = s-0007
# text = เพราะในยามน้ำท่วมสิ่งเหล่านี้จะหายากมาก
1   เพราะ   _   SCONJ   SCONJ   _   3   X   _   SpaceAfter=No
2   ใน  _   ADP ADP _   3   X   _   SpaceAfter=No
3   ยาม _   NOUN    NOUN    NounType=Class  8   X   _   SpaceAfter=No
4   น้ำท่วม   _   NOUN    NOUN    _   3   X   _   SpaceAfter=No
5   สิ่ง  _   NOUN    NOUN    NounType=Class  8   X   _   SpaceAfter=No
6   เหล่านี้   _   DET DET _   5   X   _   SpaceAfter=No
7   จะ  _   AUX AUX _   8   X   _   SpaceAfter=No
8   หา  _   VERB    VERB    _   0   ROOT    _   SpaceAfter=No
9   ยาก _   ADV ADV _   8   X   _   SpaceAfter=No
10  มาก _   ADV ADV _   9   X   _   SpaceAfter=No

# sent id = s-0008
# text = แต่ก็ไม่ควรตื่นตระหนกจนเกินไป
1   แต่  _   CCONJ   CCONJ   _   5   X   _   SpaceAfter=No
2   ก็   _   SCONJ   SCONJ   _   5   X   _   SpaceAfter=No
3   ไม่  _   PART    PART    PartType=Neg    5   X   _   SpaceAfter=No
4   ควร _   AUX AUX _   5   X   _   SpaceAfter=No
5   ตื่น  _   VERB    VERB    _   0   ROOT    _   SpaceAfter=No
6   ตระหนก   _   VERB    VERB    _   5   X   _   SpaceAfter=No
7   จน  _   SCONJ   SCONJ   _   8   X   _   SpaceAfter=No
8   เกินไป   _   ADV ADV _   5   X   _   SpaceAfter=No

# sent id = s-0009
# text = ควรมีสติต่อเหตุการณ์ที่จะเกิดขึ้น
```

NLP shared task

Multilingual Parsing from Raw Text to Universal Dependencies 2017 (http://universaldependencies.org/conll17/)

PARSEME shared task on verbal MWE identification 2017 (http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/)

SemEval shared tasks (http://alt.qcri.org/semeval2018/index.php?id=tasks)

Datasets for Natural Language Processing

https://machinelearningmastery.com/datasets-natural-language-processing/

“it is also helpful to use standard datasets that are well understood and widely used so that you can compare your results”

Text Classification

Language Modeling

Image Captioning

Machine Translation

Question Answering

Speech Recognition

Document Summarization

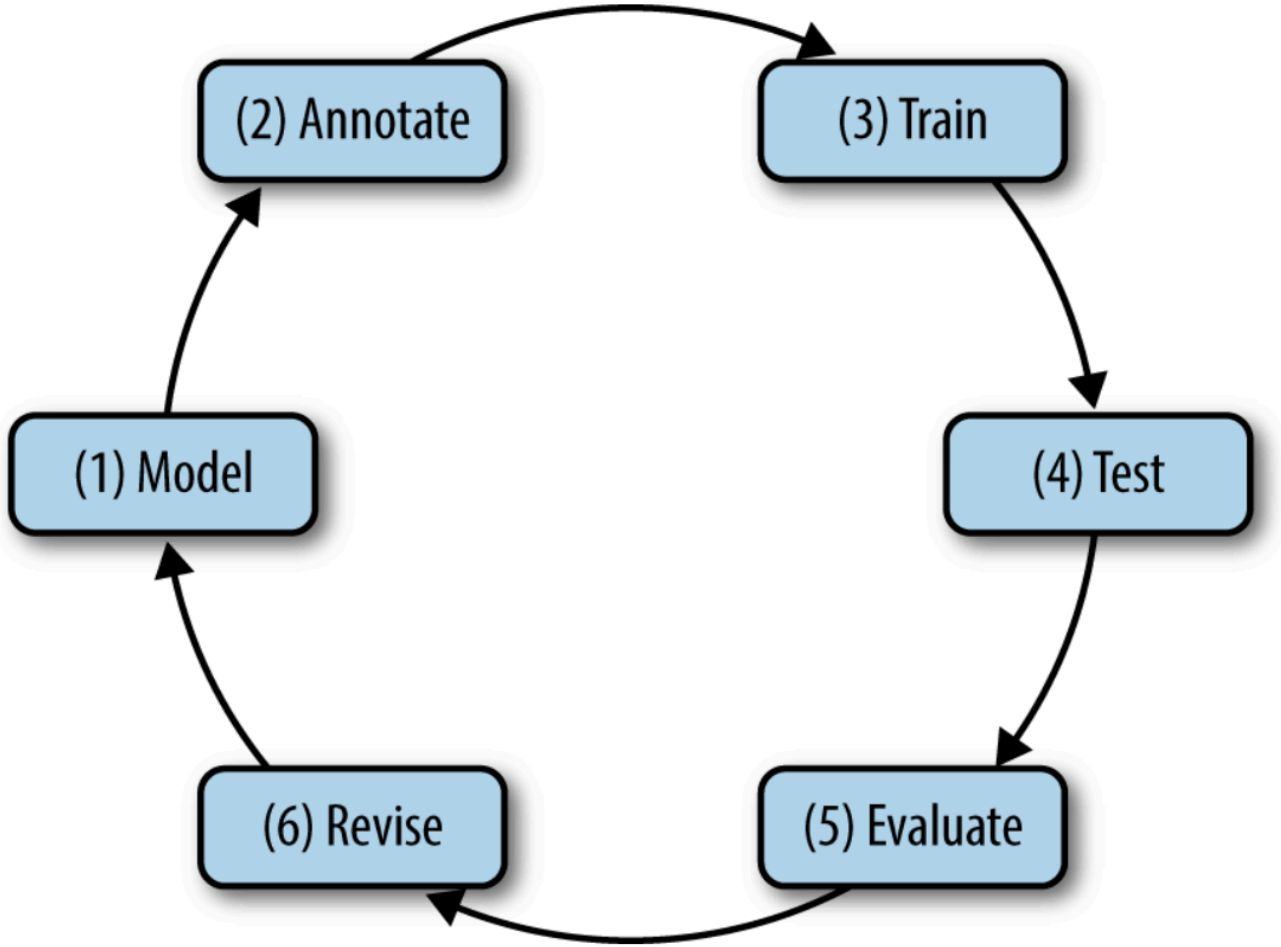


Figure 1-10. The MATTER cycle

https://www.safaribooksonline.com/library/view/natural-language-annotation/9781449332693/ch01.html

Do we need annotation?

unsupervised :

no need to annotation

testing corpus :

need correct answers

Linguistic annotation for Thai

word

multiword expression ?

named entities

sentence / utterance / elementary discourse unit

syntactic : universal dependncies

How difficult is annotation

What is a word?

ขนมไหว้พระจันทร์, ตู้เก็บกับข้าว,

เช่นเดียวกัน, นับวัน, เนื่องมาจาก

ที่จอดรถ, สถานที่จอดรถ

โรงรถ, โรงเก็บรถ, โรงเก็บรักษารถ, โรงซ่อมบำรุงรถ, โรงซ่อมบำรุงรถไฟฟ้า

How difficult is annotation

What is a POS?

which POS tag set?

ทำเพียงแค่นี้ก็พอแล้ว เพียง =? แค่ =?

ซึ่งล้วนแล้วแต่สามารถค้นหาข้อมูล ล้วน=? แล้วแต่=?

โปรแกรมบัญชีและอีกมากมาย x-และ-x อีกมากมาย=np?

ต้องไม่ยึดติดอยู่กับความอยากได้ อยากมี อยากเป็น สารพัดอยากที่ไร้ขีดจำกัด

มีอาชีพที่หลากหลาย =verb หลากหลายอาชีพ =adv อาชีพหลากหลาย =adj

How difficult is annotation

กว่าถั่วจะสุก งาก็ไหม้ กว่า=sconj

นั่งทำงานที่บ้านดีกว่านั่งทำงานที่มหาลัย กว่า=?

มีกว่าสามสิบตัว กว่า=?

ตั้งแต่เด็กจนโต เด็ก=verb? โต=noun?
ตั้งแต่เช้าจนค่ำ ตั้งแต่เล็กจนโต

จะเหบียบคันเร่งหรือเบรคดี ดี=part in this context=Q

คนรวยระดับพันล้านก็ใช่ว่าจะมีความสุข ใช่ว่า=negation = ไม่ใช่ว่า

How difficult is annotation

What is a syntactic structure?

Phrase structure tree or Dependency

Universal dependedncies

Universal POS tags

These tags mark the core part-of-speech categories. To distinguish additional lexical and grammatical properties of words, use the universal features.

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

How do we evaluate?

Thai word segmentation

- character based
- syllable based
- word based
- short word / long word รถ+โดยสาร / รถโดยสาร
- gold standard for test set

Evaluate word segmentation

character based :

char x is a segment boundary?
(x, “y”) or (x, “n”)

สาร|กึ่ง|ตัว|นำ|ที่|มี|คุณสมบัติ|ทาง|ไฟฟ้า|อยู่|ระหว่าง|ตัว|นำ|ไฟฟ้า| |และ|ฉนวน|ไฟฟ้า| |จึง|เป็น|สาร|ที่|เรา|สามารถ|ควบคุม|คุณสมบัติ|นำ|ไฟฟ้า|ของ|มัน|ได้|

remove all segment markers
=> สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และฉนวนไฟฟ้า จึงเป็นสารที่เราสามารถควบคุมคุณสมบัตินำไฟฟ้าของมันได้
Accuracy = 0.7207105914099393

original segments + add wrong segments
=>สาร|กึ่ง|ตัว|นำ|ที่|มี|คุณ^สมบัติ|ทาง|ไฟ^ฟ้า|อยู่|ระ^หว่าง|ตัว|นำ|ไฟ^ฟ้า| |และ|ฉนวน|ไฟ^ฟ้า| |จึง|เป็น|สาร|ที่|เรา|สา^มารถ|ควบคุม|คุณสมบัติ|นำ|ไฟ^ฟ้า|ของ|มัน|ได้|
Accuracy = 0.958291174817313 1

Gold standard test

wrong segment, same meaning
รถโดยสาร, เครื่องใช้ไฟฟ้า

wrong segment, no meaning/marginal meaning?
เกี่ยวกับไม้ดอกไม้ประดับ (same syllables)
พายุงวงช้างถล่ม, (different syllables)

different segment, different meaning (depend on context)
ทางการเมือง, ที่อยู่อาศัย, น้ำท่วม (same syllables)
ตากลม, หลวงตามหาบัว (different syllables)

unknown words