

# 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 Initiative

← → ↻ ⓘ www.tei-c.org/index.xml ☆ 📁 📄 🤖 📖 📌 ⋮

📱 Apps 📁 + Pocket 📁 Arts 📁 Digital Humanities 📧 Mail CU 📧 Gmail 📘 Facebook 📺 SCBEasy 📌 Bualuang 🌐 K-Cyber 📄 User Ident 🏠 Me » 📁 Other Bookmarks

 <Text Encoding Initiative>

[Home](#) [Guidelines](#) [Activities](#) [Tools](#) [Membership](#) [Support](#) [About](#) [News](#)

[Home](#)  Entire site 🔍 **Search**

## TEI-C News

[Call for Papers – TEI 2018 Tokyo, Japan](#)

*Posted on: 2018-03-06*

[TEI Guidelines – Version 3.3.0](#)

*Posted on: 2018-01-31*

[TEI-C and TAPAS Call for Nominations 2017](#)

*Posted on: 2017-08-24*

[TEI Guidelines – Version 3.2.0](#)

*Posted on: 2017-07-11*

[submissions open for 2017 Members Meeting and Conference](#)

*Posted on: 2017-04-09*

[Call for Nominations: Rahtz Prize for TEI Ingenuity](#)

*Posted on: 2017-03-31*

## Other News

[2018 and 2019 TEI Conference and Members' Meeting](#)

*Posted on: 2017-09-11*

[TEI Consortium/TEI Community Awarded ADHO's Antonio Zampolli Prize!](#)

*Posted on: 2016-07-19*

[2016 TEI conference: programme published, early](#)

## TEI: Text Encoding Initiative

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of [resources](#) and [training events](#) for learning TEI, information on [projects using the TEI](#), a [bibliography of TEI-related publications](#), and [software](#) developed for or adapted to the TEI.

The TEI Consortium is a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors. In commemoration of the TEI community's 30th anniversary, it will be awarded the 2017 Antonio Zampolli Prize from the Alliance of Digital Humanities Organizations.

Want to become active in the TEI community? [Become a TEI Member](#), join a [special interest group](#), sign up for the [TEI-L mailing list](#), and come to our [annual conferences and members' meetings](#).

*Last recorded change to this page: 2016-07-19 • For corrections or updates, contact webmaster AT tei-c.org*

## Front Matter

### Title

- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- ⊕ iv. [About These Guidelines](#)
- ⊕ v. [A Gentle Introduction to XML](#)
- ⊕ vi. [Languages and Character Sets](#)

## Back Matter

- ⊕ Appendix A [Model Classes](#)
- ⊕ Appendix B [Attribute Classes](#)
- ⊕ Appendix C [Elements](#)
- ⊕ Appendix D [Attributes](#)
- ⊕ Appendix E [Datatypes and Other Macros](#)
- ⊕ Appendix F [Bibliography](#)
- ⊕ Appendix G [Prefatory Notes](#)
- Appendix H [Colophon](#)

## Text Body

- ⊕ 1 [The TEI Infrastructure](#)
- ⊕ 2 [The TEI Header](#)
- ⊕ 3 [Elements Available in All TEI Documents](#)
- ⊕ 4 [Default Text Structure](#)
- ⊕ 5 [Characters, Glyphs, and Writing Modes](#)
- ⊕ 6 [Verse](#)
- ⊕ 7 [Performance Texts](#)
- ⊕ 8 [Transcriptions of Speech](#)
- ⊕ 9 [Dictionaries](#)
- ⊕ 10 [Manuscript Description](#)
- ⊕ 11 [Representation of Primary Sources](#)
- ⊕ 12 [Critical Apparatus](#)
- ⊕ 13 [Names, Dates, People, and Places](#)
- ⊕ 14 [Tables, Formulae, Graphics and Notated Music](#)
- ⊕ 15 [Language Corpora](#)
- ⊕ 16 [Linking, Segmentation, and Alignment](#)
- ⊕ 17 [Simple Analytic Mechanisms](#)
- ⊕ 18 [Feature Structures](#)
- ⊕ 19 [Graphs, Networks, and Trees](#)
- ⊕ 20 [Non-hierarchical Structures](#)
- ⊕ 21 [Certainty, Precision, and Responsibility](#)
- ⊕ 22 [Documentation Elements](#)
- ⊕ 23 [Using the TEI](#)

## 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

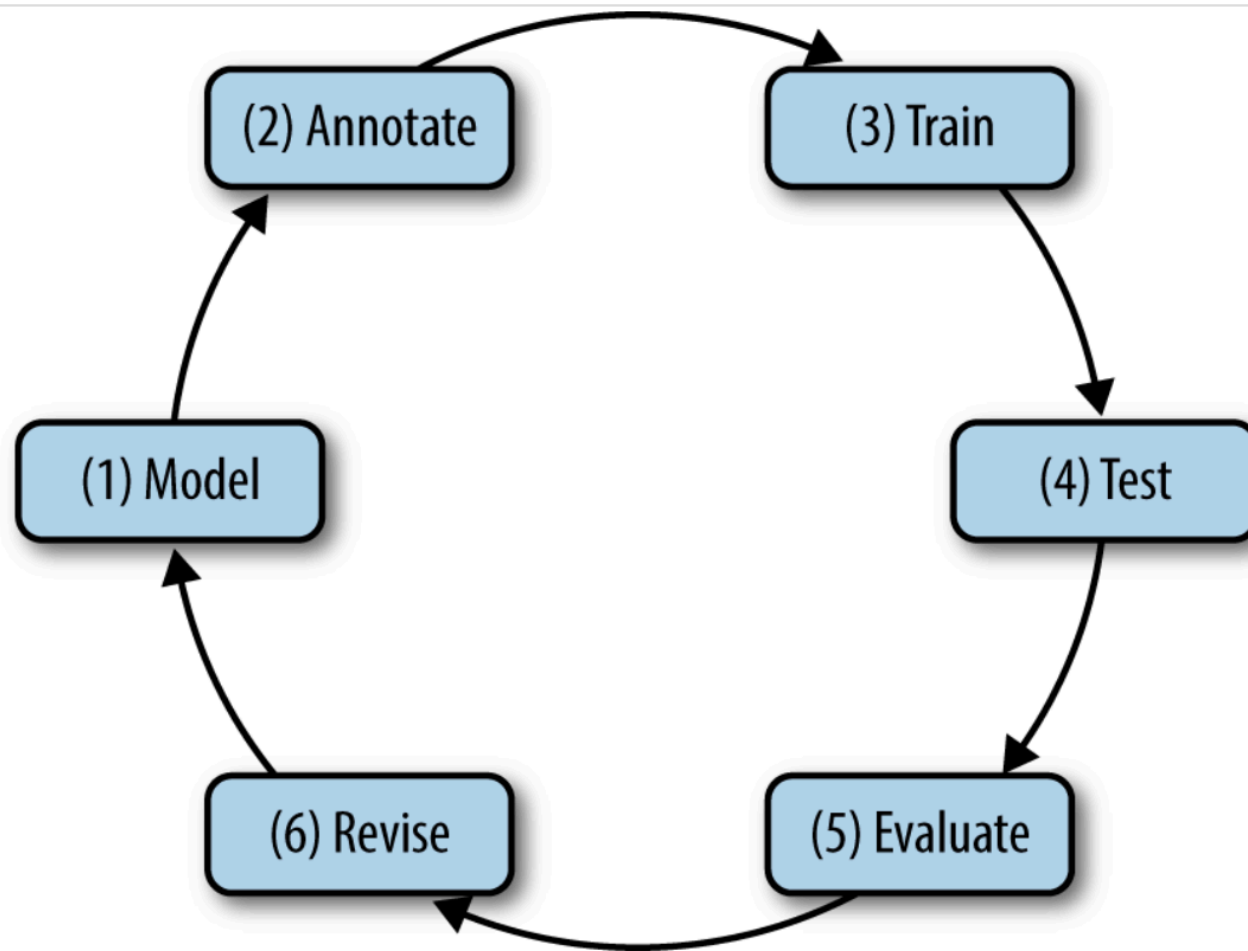
Machine Translation

Question Answering

Speech Recognition

Document Summarization





*Figure 1-10. The MATTER cycle*

# 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 dependencies

# How difficult is annotation

What is a word?

ขนมไหว้พระจันทร์, ตู้เก็บกับข้าว,

เช่นเดียวกัน, นับวัน, เนื่องจาก

ที่จอดรถ, สถานที่จอดรถ

โรงรถ, โรงเก็บรถ, โรงเก็บรักษารถ, โรงซ่อมบำรุงรถ, โรง  
ซ่อมบำรุงรถไฟฟ้า

# How difficult is annotation

What is a POS?

which POS tag set?

ทำเพียงแค่นี้ก็พอแล้ว เพียง=? แค่นี้=?

ซึ่งล้วนแล้วแต่สามารถค้นหาข้อมูล ล้วน=? แล้วแต่=?

โปรแกรมบัญชีและอีกมากมาย x-และ-x อีกมากมาย=np?

ต้องไม่ยึดติดอยู่กับความอยากได้ อยากมี อยากเป็น สารพัดอยากที่ไร้ขีดจำกัด

มีอาชีพที่หลากหลาย =verb หลากหลายอาชีพ =adv อาชีพหลากหลาย =adj

# How difficult is annotation

กว่าถั่วจะสุก งามก็ไหม้ กว่า=sconj

นั่งทำงานที่บ้านดีกว่านั่งทำงานที่มหาลัย กว่า=?

มีกว่าสามสิบตัว กว่า=?

ตั้งแต่เด็กจนโต เด็ก=verb? โต=noun?

ตั้งแต่เข้าจนคำ ตั้งแต่เลิกจนโต

จะเหยียบคันเร่งหรือเบรคดี ดี=part in this context=Q

คนรวยระดับพันล้านก็ ใช่ว่าจะมีความสุข ใช่ว่า=negation = ไม่ใช่ว่า

# How difficult is annotation

What is a syntactic structure?

Phrase structure tree or Dependency

Universal dependencies

# 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
<a href="#">ADJ</a>	<a href="#">ADP</a>	<a href="#">PUNCT</a>
<a href="#">ADV</a>	<a href="#">AUX</a>	<a href="#">SYM</a>
<a href="#">INTJ</a>	<a href="#">CCONJ</a>	<a href="#">X</a>
<a href="#">NOUN</a>	<a href="#">DET</a>	
<a href="#">PROPN</a>	<a href="#">NUM</a>	
<a href="#">VERB</a>	<a href="#">PART</a>	
	<a href="#">PRON</a>	
	<a href="#">SCONJ</a>	

---



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

# 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|คุณ|สมบัติ|ทาง|ไฟ|ฟ้า|อยู่|ระ|หว่า|ง|ตัว|นำ|ไฟ|ฟ้า| |แ|ล|และ|ณ|นวน|ไฟ|ฟ้า| |จ|ึง|เป็น|สาร|ที่|เร|า|  
สา|มารถ|คว|บ|ค|คุณ|สมบัติ|นำ|ไฟ|ฟ้า|ของ|มัน|ได้|

remove all segment markers

=> สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และนวนไฟฟ้า จึงเป็นสารที่เราสามารถ  
ควบคุมคุณสมบัตินำไฟฟ้าของมันได้

Accuracy = 0.7207105914099393

original segments + add wrong segments

=> สาร|กึ่ง|ตัว|นำ|ที่|มี|คุณ^|สมบัติ|ทาง|ไฟ^|ฟ้า|อยู่|ระ^|หว่า|ง|ตัว|นำ|ไฟ^|ฟ้า| |แ|ล|และ|ณ|นวน|ไฟ^|ฟ้า| |จ|ึง|เป็น|  
สาร|ที่|เร|า|สา^|มารถ|คว|บ|ค|คุณ|สมบัติ|นำ|ไฟ^|ฟ้า|ของ|มัน|ได้|

Accuracy = 0.9582911748173131

# Gold standard test

wrong segment, same meaning

รถโดยสาร, เครื่องใช้ไฟฟ้า

wrong segment, no meaning/marginal meaning?

เกี่ยวกับไม้ดอกไม้ประดับ (same syllables)

พายุวงช้างถล่ม, (different syllables)

different segment, different meaning (depend on context)

ทางการเมือง, ที่อยู่อาศัย, น้ำท่วม (same syllables)

ตากลม, หลวงตามหาบัว (different syllables)

unknown words