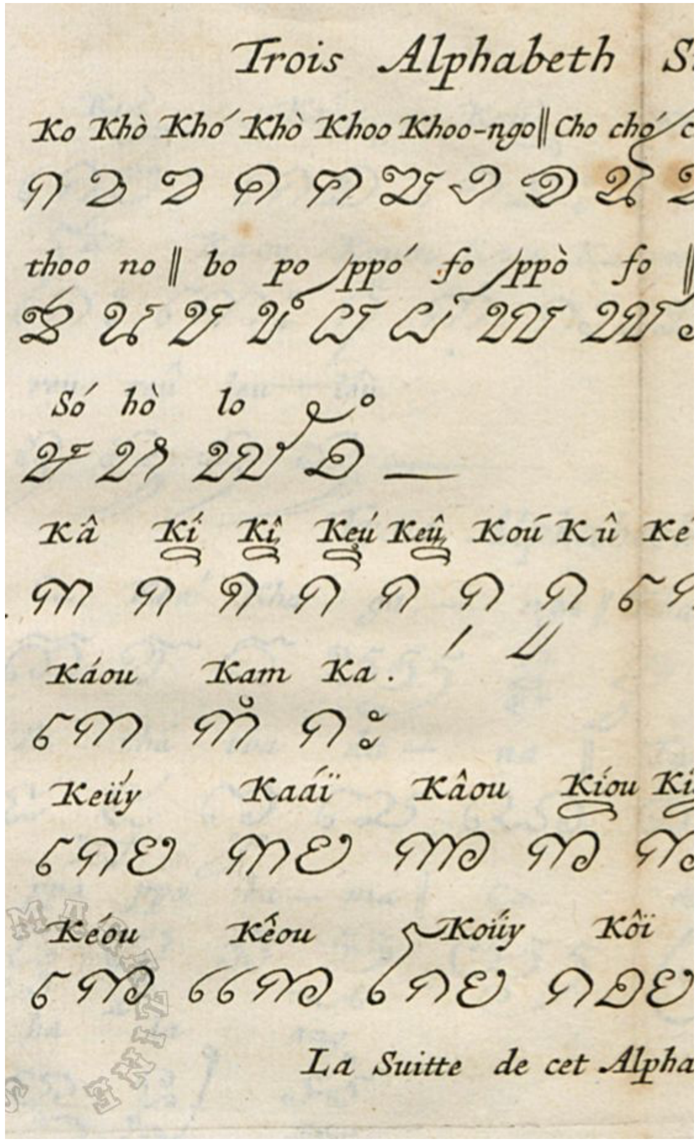
A Brief Specification of LST20 Corpus

Prachya Boonkwan, Vorapon Luantangsrisuk, Sitthaa Phaholphinyo, Kanyanat Kriengket, Dhanon Leenoi, Charun Phrombut, Monthika Boriboon, Krit Kosawat, and Thepchai Supnithi

Language and Semantic Technology Lab (LST)
National Electronics and Computer Technology Center (NECTEC)
Bangkok, Thailand



Outline

- Executive summary
- Agreement of usage & download
- Annotation schemes
- Statistics

NECTEC
a member of NSTDA

Executive Summary

Executive Summary

NECTEC
a member of NSTDA

- LST20 Corpus
  - Dataset for training fundamental Thai language processing tasks
  - Featured linguistic information
    - **Word boundaries** for word segmentation
    - **Named entities** for named entity recognition
    - **Clause boundaries** for clause segmentation
    - **Sentence boundaries** for sentence segmentation
    - **News genres** for document classification
  - CoNLL-2003 format: tab-separated columns

Words	3,163,034
Named entities	288,020
Clauses	248,181
Sentences	74,180
Distinct words	46,692
Genres	15
News articles	3,745

Available at https://aiforthai.in.th

Agreement of Usage & Download

Agreement of Usage

- Data consortium
  - Non-commercial use, research, open source
    - Free of charge
    - Citing our technical paper is always required
  - Commercial use
    - **Option 1 (in-kind):** Contributing a dataset of 150K-300K words completely annotated with our annotation schemes within 1 year
    - **Option 2 (in-cash):** Purchasing the dataset

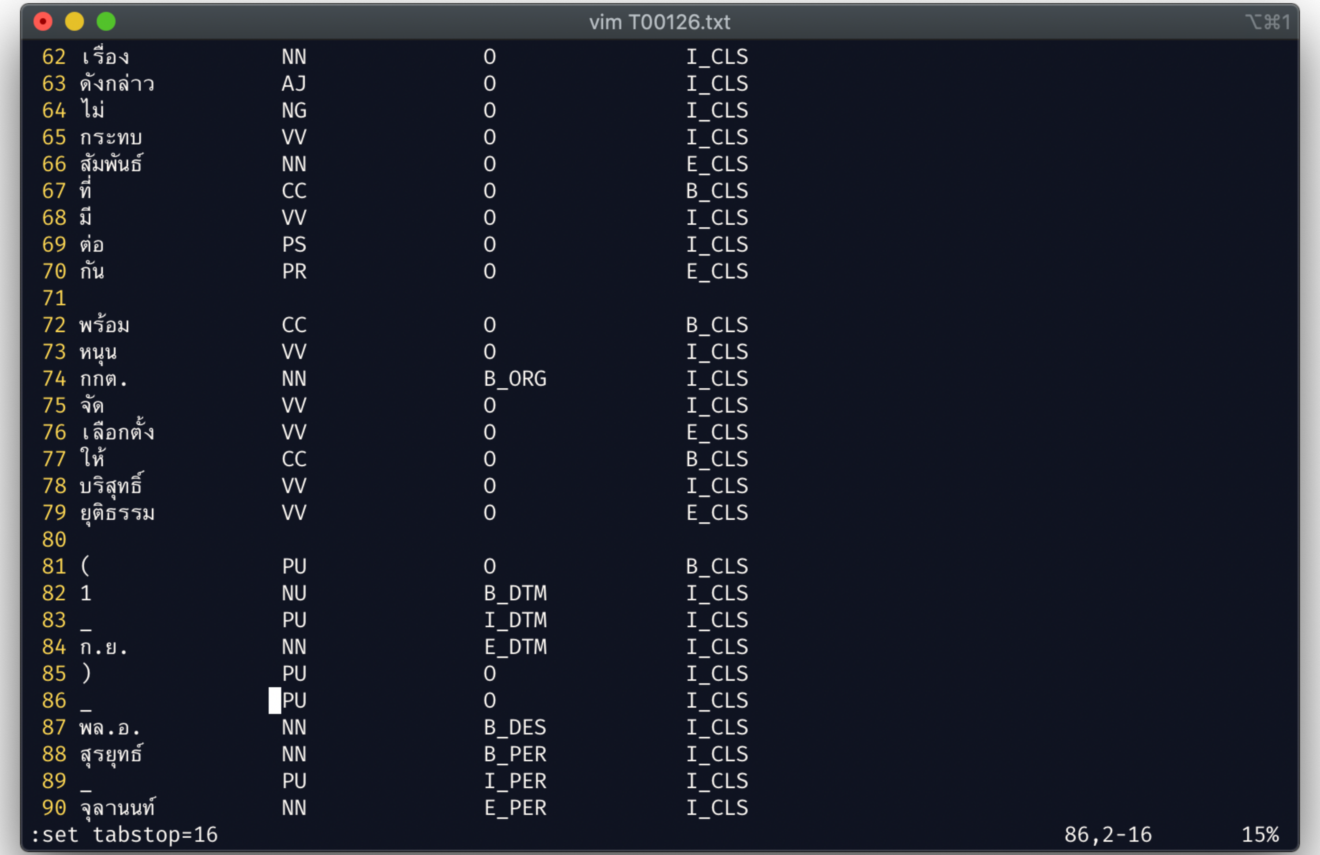
AI FOR THAI

- **Download**
  - https://aiforthai.in.th
  - Registration is required before downloading
  - Much more web APIs and demos are also available therein

NECTEC
a member of NSTDA

Annotation Schemes

Format: CoNLL-2003 Style



```
เรื่อง           NN              O               I_CLS
ดังกล่าว          AJ              O               I_CLS
ไม่             NG              O               I_CLS
กระทบ           VV              O               I_CLS
สัมพันธ์          NN              O               E_CLS
ที่              CC              O               B_CLS
มี              VV              O               I_CLS
ต่อ             PS              O               I_CLS
กัน             PR              O               E_CLS

พร้อม           CC              O               B_CLS
หนุน            VV              O               I_CLS
กกต.            NN              B_ORG           I_CLS
จัด             VV              O               I_CLS
เลือกตั้ง          VV              O               E_CLS
ให้             CC              O               B_CLS
บริสุทธิ์          VV              O               I_CLS
ยุติธรรม          VV              O               E_CLS

(               PU              O               B_CLS
1               NU              B_DTM           I_CLS
_               PU              I_DTM           I_CLS
ก.ย.            NN              E_DTM           I_CLS
)               PU              O               I_CLS
_               PU              O               I_CLS
พล.อ.           NN              B_DES           I_CLS
สุรยุทธ์          NN              B_PER           I_CLS
_               PU              I_PER           I_CLS
จุลานนท์          NN              E_PER           I_CLS
:set tabstop=16                                                   86,2-16       15%
```

- Four columns
  - Word
  - POS tag
  - Named entity
  - Clause boundary
- Notes
  - Each column is separated by a tab
  - Empty line marks sentence boundary

POS Tagset

NECTEC
a member of NSTDA

Tags	Names	Brief Descriptions
AJ	Adjective	Attribute, modifier, or description of a noun
AV	Adverb	Word that modifies or qualifies an adjective, verb, or another adverb
AX	Auxiliary	Tense, aspect, mood, and voice
CC	Connector	Conjunction and relative pronoun
CL	Classifier	Class or measurement unit to which a noun or an action belongs
FX	Prefix	Inflectional (nominalizer, adjectivizer, adverbializer, and courteous verbalizer), and derivational
IJ	Interjection	Exclamation word
NG	Negator	Word of negation

Tags	Names	Brief Descriptions
NN	Noun	Person, place, thing, abstract concept, and proper name
NU	Number	Quantity for counting and calculation
PA	Particle	Politeness, intention, belief, question
PR	Pronoun	Word used to refer to an element in the discourse
PS	Preposition	Location, comparison, instrument, exemplification
PU	Punctuation	Punctuation mark
VV	Verb	Action, state, occurrence, and word that forms the predicate part
XX	Others	Unknown category

* Green texts = content word | Black texts = function word | Red texts = undesirable (yet they still exist)

Named Entities

Tags	Names	Descriptions
TTL	Title	Family relationship, social relationship, and permanent title
DES	Designation	Position and professional title
PER	Person	Name of a person or family
ORG	Organization	Name of organization, office, or company
LOC	Location	Name of land according to geo-political borders
BRN	Brand	Name of brand, product, and trademark
DTM	Date and time	Time or a specific period of time
MEA	Measurement	Measurement unit and quantity of things
NUM	Number	The number of a measurement unit
TRM	Terminology	Domain-specific word

Green texts = personal entity | Red texts = collective entity | Black texts = referential entity

- BIEO tagging convention
  - B_tag = beginning
  - I_tag = intermediate/inside
  - E_tag = ending
  - O_tag = outside
- Notes
  - BIEO convention can be converted to BIO by replacing all E_tag's with I_tag

Clause Boundary

- **Definition:** clause = a part of sentence that contains at least one verb
- Syntactic clues for clause boundary
  - Subordinate connector e.g. ซึ่ง, ถ้า, ว่า
  - Cohesive marker e.g. อย่างไรก็ตาม, นอกจากนี้
  - List marker e.g. เช่น, ได้แก่, ตามลำดับ
  - Particle e.g. ครับ, นะ
  - Question adverbs e.g. อย่างไร, ไหม

Sentence Boundary

- **Definition:** sentence = (1) a group of at least one clause, or (2) a phrase acting as a topic
- Clues for sentence segmentation
  - Topic shift denoted with a cohesive marker: **break**
  - Subject shift between two adjacent clauses: **break**
  - Direct speech "………": **concatenate** inside quotes
  - Indirect speech: **break** when the subject changes
  - Item list: **concatenate** all items
  - Particle e.g. ครับ, นะ: **break**

Statistics

News Genres

Politics	Royal
Crime and Accident	Disaster
Economics	Urban development
Entertainment	Environment
Sports	Culture
International	Weather forecast
Science, Technology, and Education	General
Health	

- Specification
  - **Size:** 3,745 articles
  - **Sources:** Thairath, Dailynews, Manager, Matichon, Nation, Prachachat Business
  - **Duration:** JAN 2003 to DEC 2009

NECTEC
a member of NSTDA

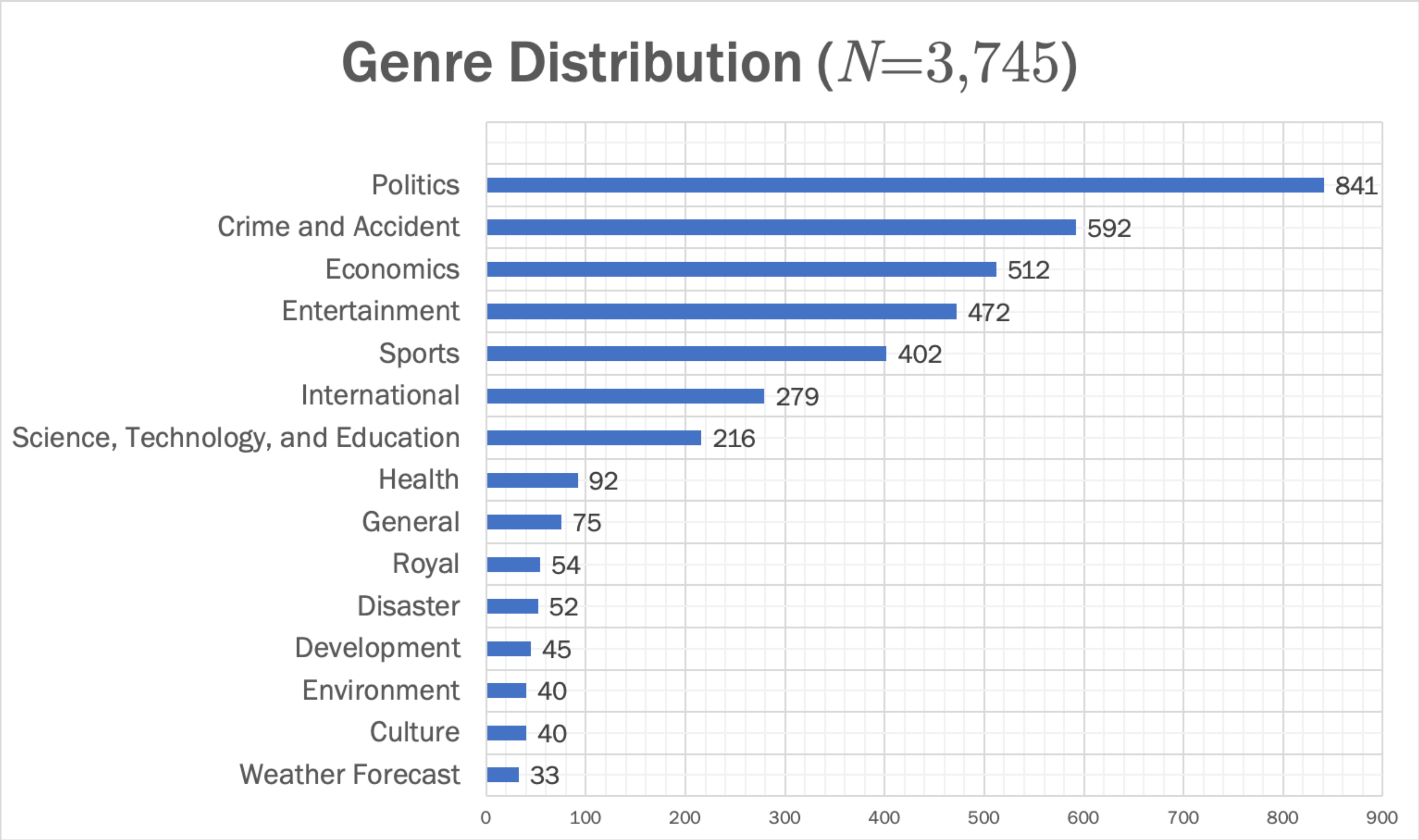
Corpus Size

	Training	Evaluation	Testing	Total
Words	2,714,848	240,891	207,295	3,163,034
Named entities	246,529	23,176	18,315	288,020
Clauses	214,645	17,486	16,050	248,181
Sentences	63,310	5,620	5,250	74,180
Distinct words	42,091	(oov) 2,595	(oov) 2,006	46,692
Breaking spaces*	63,310	5,620	5,250	74,180
Non-breaking spaces**	402,380	39,920	32,204	475,504

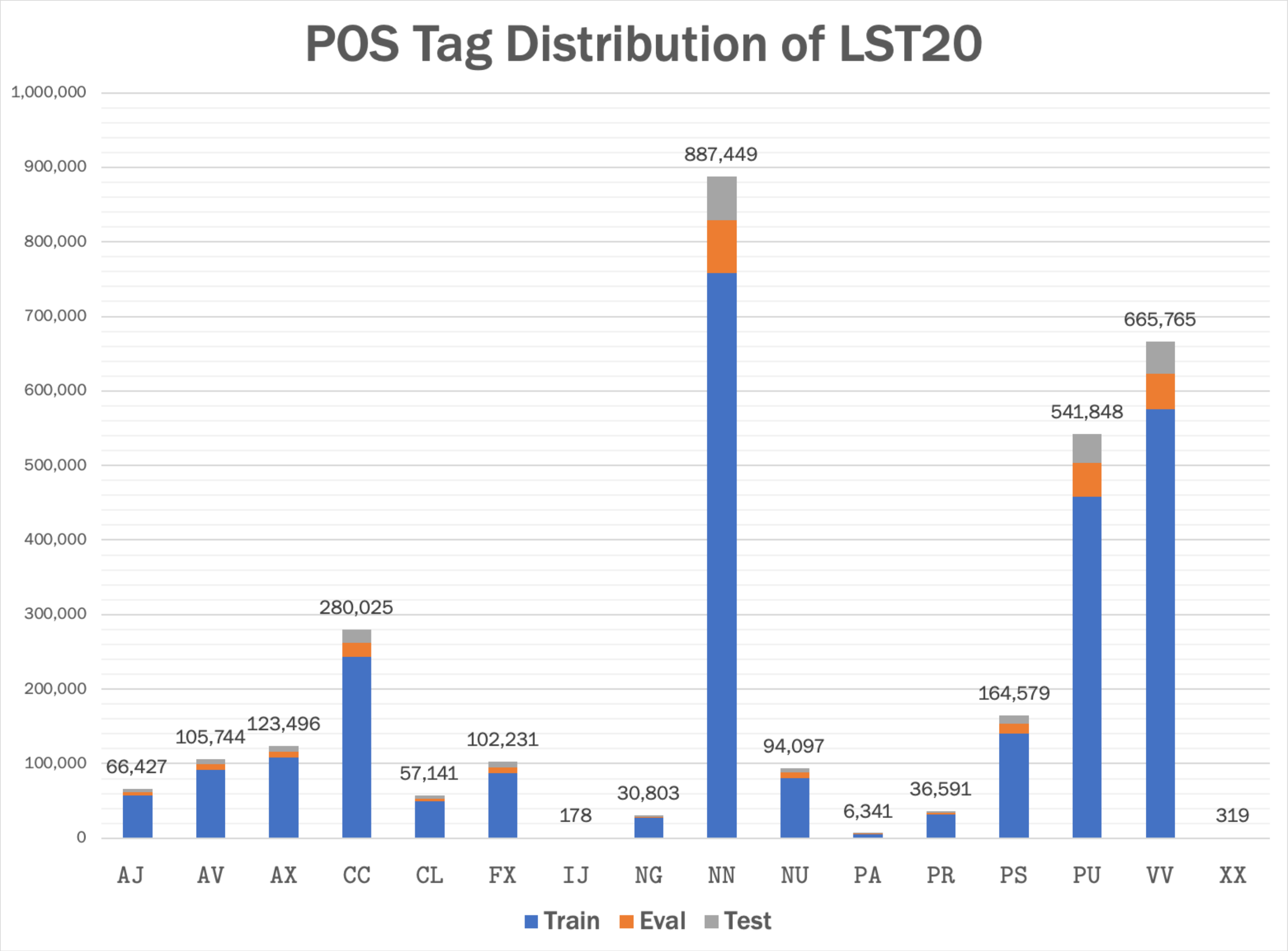
* Breaking space = space that is used as a sentence boundary marker

** Non-breaking space = space that is not used as a sentence boundary marker

Statistics: New Genres



Statistics: POS Distribution



NECTEC
a member of NSTDA

Thank You