FACULTY OF ARTS AND PHILOSOPHY

Contact

Daniel.kong@ugent.be
https://lt3.ugent.be/people/daniel-kong/

@DanielKong2021

LANGUAGE AND TRANSLATION TECHNOLOGY TEAM (LT3)

Delu KONG, Lieve Macken

DECODING MACHINE TRANSLATIONESE: ASSESSING THE PERFORMANCE OF LARGE LANGUAGE MODELS IN ENGLISH-CHINESE NEWS TRANSLATION

Background

❖ Large language models (LLMs) have driven neural machine translations (NMTs) into a new era, with news translation greatly influenced.

❖ People are exposed to an increasing number of MTed news. So it is vital to evaluate **machine translationese (MTese)** in NMTed & LLMTed news articles.

Previous studies

❖ "Translationese" has been used to refer to the characteristics that distinguish original text from translated text (Gellerstam 1986), and MTed texts may also contain such features (Daems et al. 2017).

❖ De Clercq et al. (2021) found MTed French do vary from original French in features such as sentence length.

Research questions

❖ RQ1: Do MTese exist in Eng -> Chi MTed news? Through which features?

❖ RQ2: Do LLMs differ from NMTs in terms of MTese?

language and translation technology team
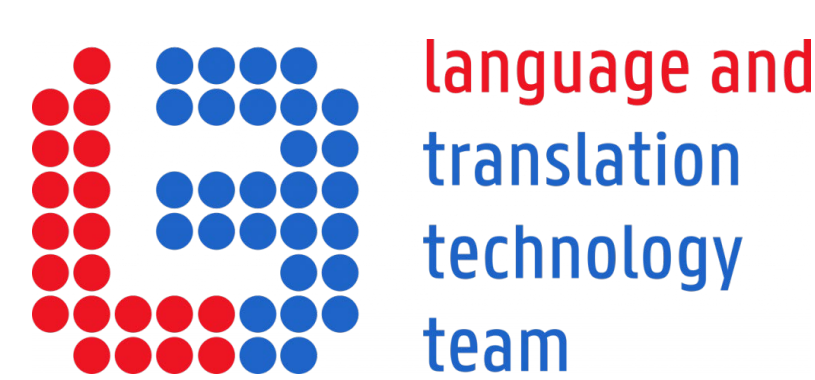
Corpus Design

4 corpora designed to represent different Original & MT text types

| Corpus	Type	Abbr	Texts	Token
OCN	Orig. Chi. News	OCN	200	186 052
OEN	Orig. Eng. News	OEN	200	312 587
NMT	Google	NGT	200	367 107
	DeepL	NDL	200	383 131
	Microsoft	NMS	200	385 208
	Baidu	NBD	200	354 721
	Youdao	NYD	200	366 658
LLM	ChatGPT(4o)	LCG	200	325 060
	Claude(3.5sonnet)	LCL	200	349 010
	Gemini(1.5flash)	LGM	189	347 300
	Kimi(v1-8k)	LKM	185	292 550
	ChatGLM(4-plus)	LGL	178	347 300
	UnbabelTower(7B)	LTO	200	335 680

❖ News selection: Original Chinese news from *People's Daily (人民日报)* and *Xinhua News(新华网)*, etc. Original English news from *The Economist*, and *The Guardian* etc.

❖ All files are **cleaned, denoised, segmented, PoS-tagged and Dep-tagged**.

❖ 5 NMT engines: 2 Chinese firms
6 LLM engines: 2 Chinese, 1 MT specific (LTO)

❖ NMTed texts are all obtained through API. LLMTed also mainly through API, while LCG & LCL from their official web.
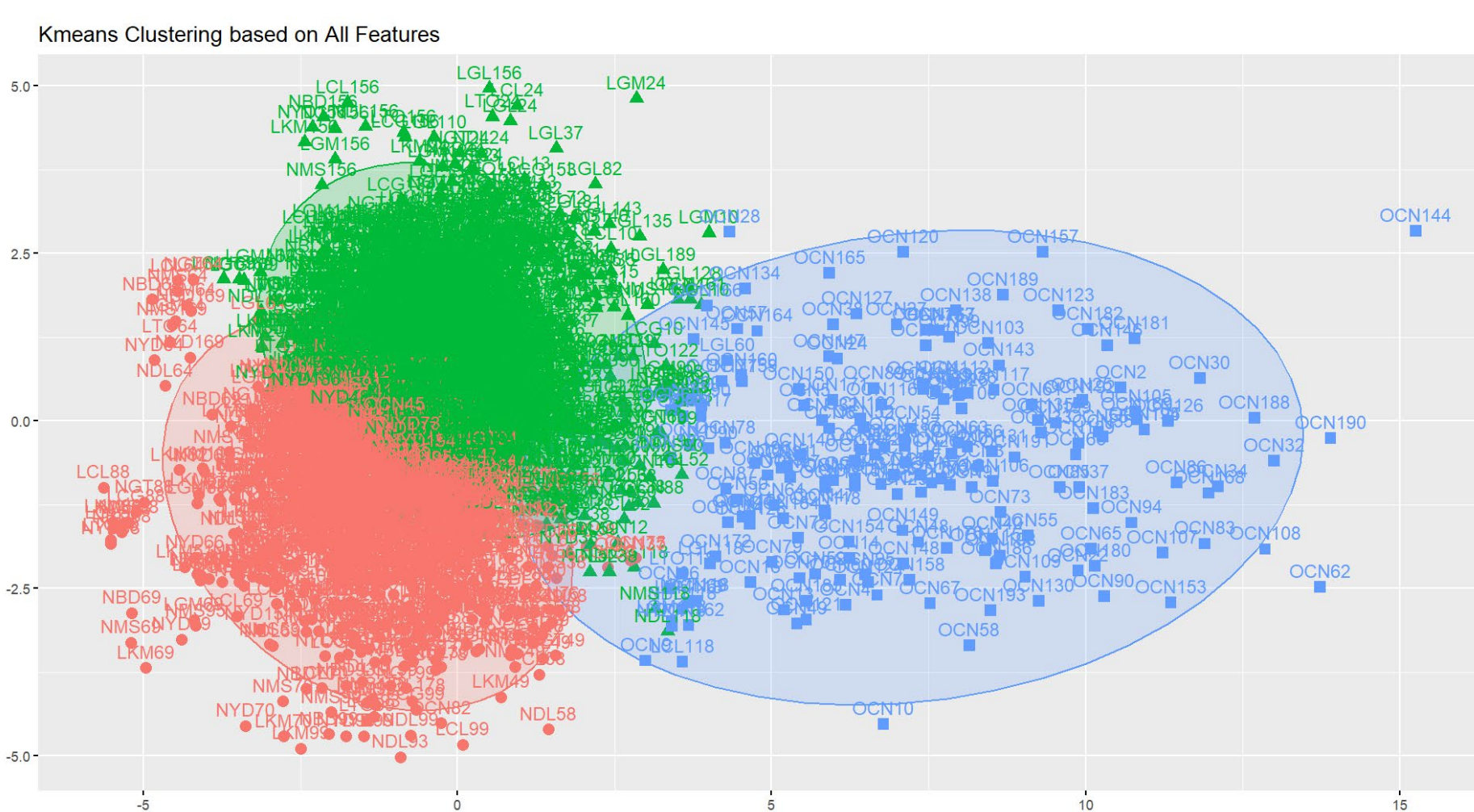
Classification & Clustering

(A) SVM classifier: best performance between OCN vs. MTs

Text A	Text B	Correct / Total Samples	Accuracy	Top 3 Features
OCN	NMTs & LLMs	4285 / 4352	97.61%	3rdPron, ContWords, Noun
OCN	LLMs	2306 / 2352	97.11%	ContWords, 3rdPron, Declaratives
OCN	NMTs	1979 / 2000	98.20%	3rdPron, ContWords, Declaratives
NMTs	LLMs	9882 / 11760	82.02%	Words per sentence, Brackets, comma
NMTs	NMTs	3426 / 4000	83.93%	Words per sentence, Brackets, progConj
LLMs	LLMs	4588 / 5760	77.09%	comma, Declaratives, Words per sentence

❖ OCN are different from all other MT engines, both NMTs and LLMs. (bottom red cells)

❖ **(1)** shows classification within LLMs are less accurate than others.

❖ **(2)** shows LLMs and NMTs are quite different except NBD & NYD.

❖ **(3)** shows NMTs texts are quite similar except NBD & NYD.

❖ RQ1 answered: **Compared with ONC, MTese do exists in both NMTs and LLMs.**

Pairwise Classification Accuracy Heatmap
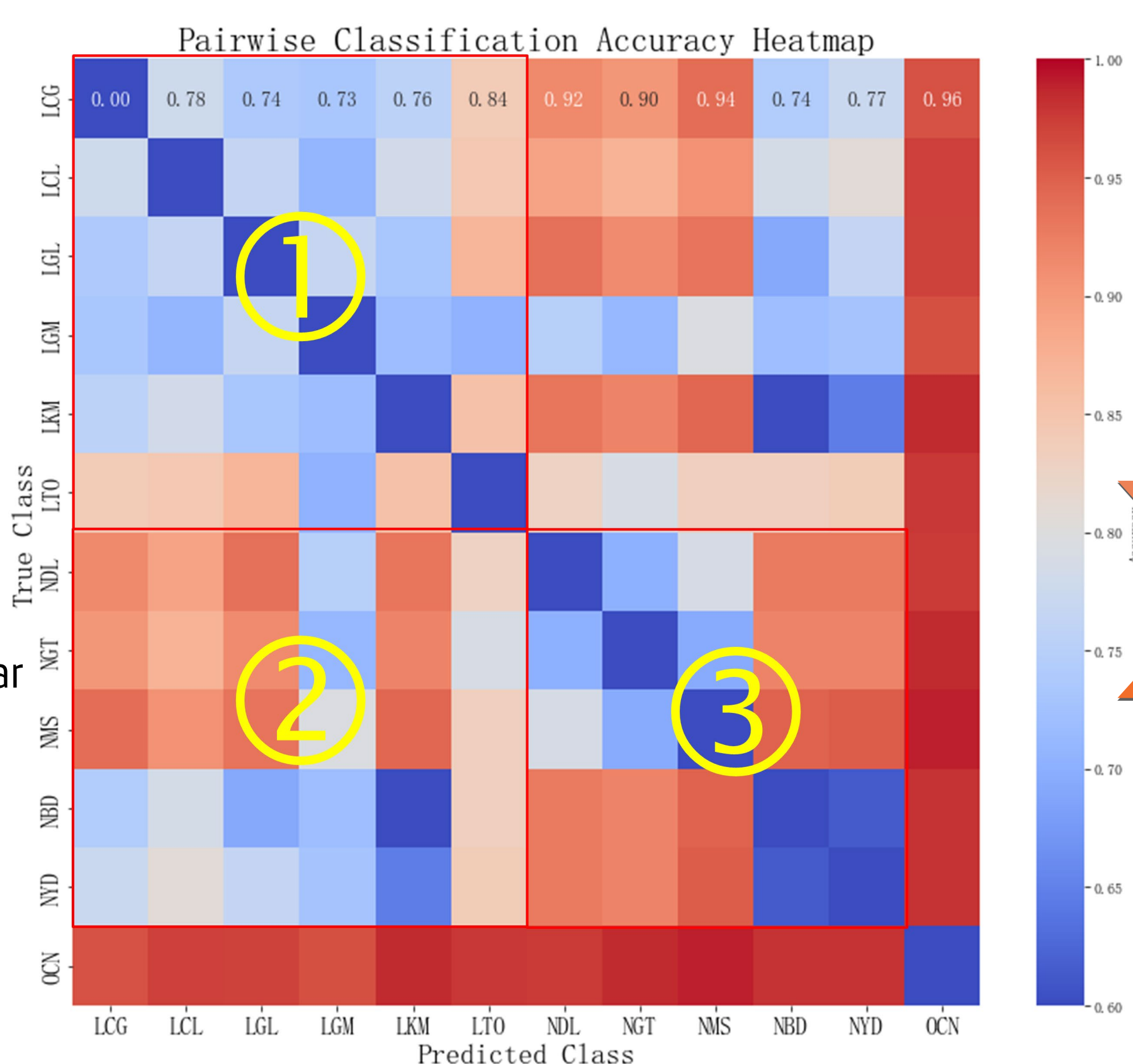
(B) Kmeans Clustering: Confused distribution between LLMs & NMTs

Kmeans Clustering based on All Features

- OCN (blue squares) are clustered together, different from MTs.

- RQ2: **NMTs and LLMs are partially mixed, meaning LLMs are still sharing MTese features with NMTs. While LLMs are also partially separated from NMTs, so certain MTese features are quite different.**

Feature set

Vectorizing the corpora in question using 97 linguistic features

Feature Set

Lexical level — Syntactical level

General lexical features | PoS-tags-based features | General syntax features | Dep-tags-based features

General lexical features: TTR, STTR, Word length, Hapax Legomena — Total 15 features

PoS-tags-based features: Noun ratio, Verb ratio, Content words, Speical words — Total 52 features

General syntax features: Sentence length, Punctuation ratio, Sentence patterns, Hapax Legomena — Total 10 features

Dep-tags-based features: MDD, Dep. relations, Avg. Tree height, Leaf node ratio — Total 20 features

All together 97 features.

Illustration

ANOVA analysis of 4 distinctive features' distribution among texts

ANOVA Analysis of 3rdpron Frequency — ANOVA F-statistic: 102.00, p-value: 1.22e-206 (A)

ANOVA Analysis of Words Per Sentence — ANOVA F-statistic: 165.22, p-value: 3.26e-307 (B)

ANOVA Analysis of Bracket Frequency — ANOVA F-statistic: 86.64, p-value: 3.41e-179 (C)

ANOVA Analysis of Comma Frequency — ANOVA F-statistic: 22.16, p-value: 2.66e-47 (D)

❖ **A** : 3rd pronoun: OEN > NMTs ≈ LLMs > OCN

❖ **B** : Words per sent: OCN > NMTs ≈ LLMs > OEN

❖ **C** : Bracket: NMTs > LLMs ≈ OEN > OCN

❖ **D** : Comma: OCN ≈ LLMs > NMTs > OEN

LLMs differentiated from NMTs in subtle features like puncs ratio

Future work: To do more feature engineering to distinguish LLMs and NMTs satisfactorily; To investigate whether language-preference or domain-specific engines may show difference when compared.
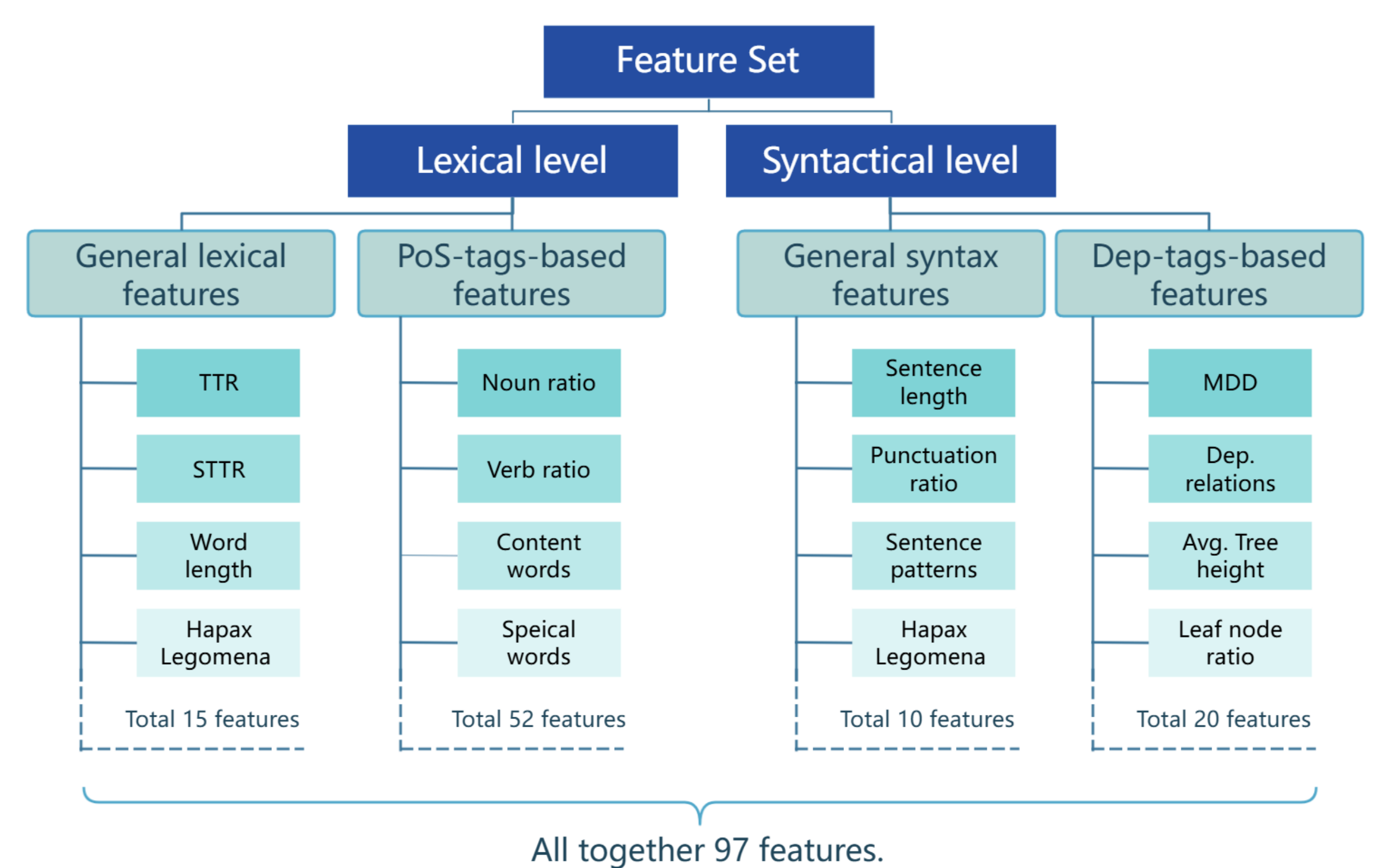
Reference

Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies, 16*, 89–103. https://doi.org/10.52034/lanstts.v16i0.434

De Clercq, O., De Sutter, G., Loock, R., Cappelle, B., & Plevoets, K. (2021). Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French. *Translation Quarterly, 101*, 21–45.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia, 1*, 88–95.
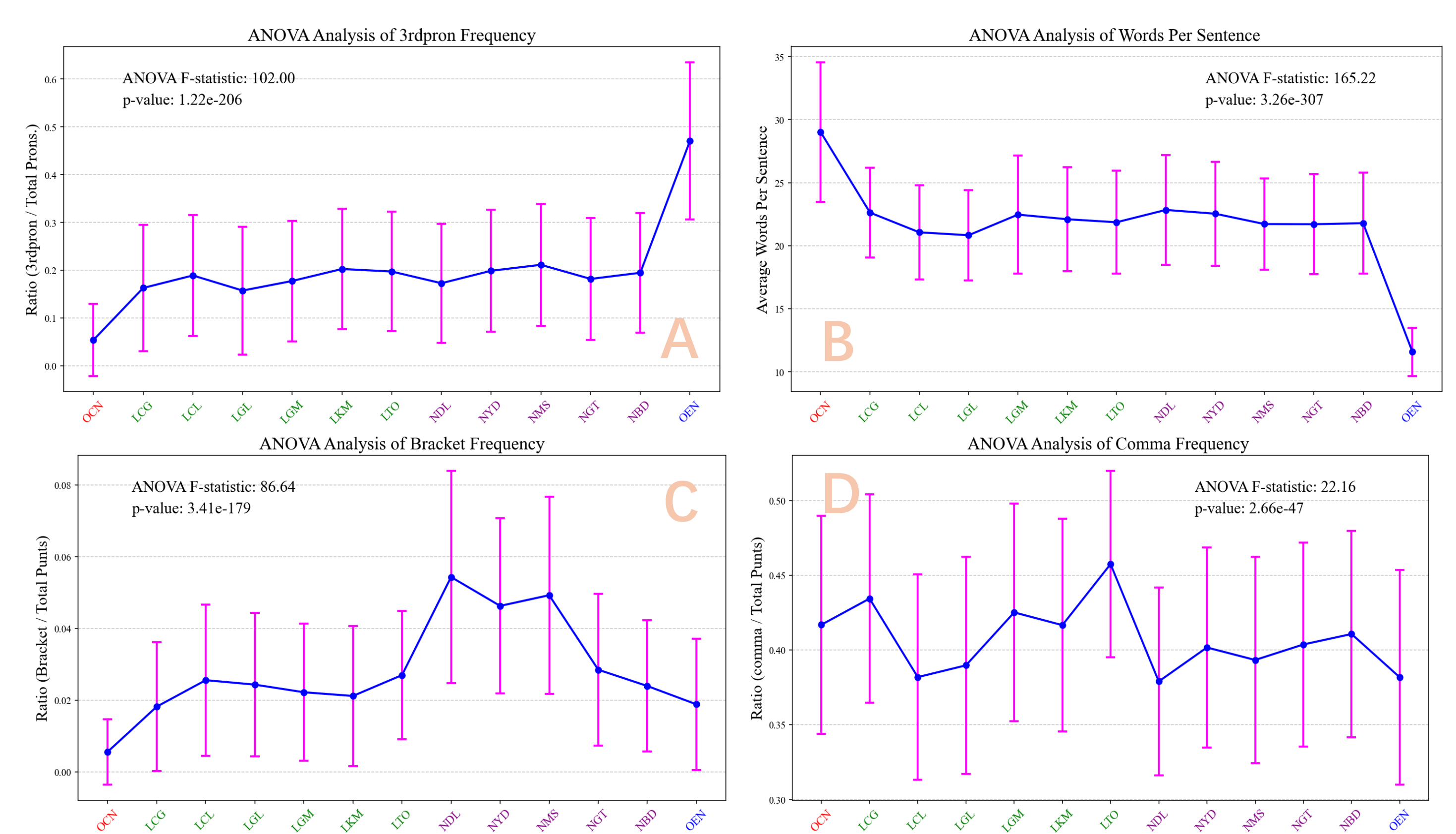
GHENT UNIVERSITY