Application of Information Theory in Translation(ese) Studies

Maria Kunilovskaya

April 12, 2024

Outline

1. Translation(ese) Studies
2. Enter Information Theory
   - Previous work
   - B7 research designs and findings
3. Current research
   - Monolingual models from parallel data
   - In quest for the best modelling approach
4. References

Translationese studies

Translationese studies is a central direction in Translation Studies that aims to capture and explain the specificity of translations, often with regard to:

- translation direction (distance between SL and TL): e.g. EN$\rightarrow$DE$\neq$DE$\rightarrow$EN
- register: e.g. translationese in fiction vs. academic writing
- level of expertise: professional vs student
- mode: translation vs interpreting

Translation (interpreting) and language variation

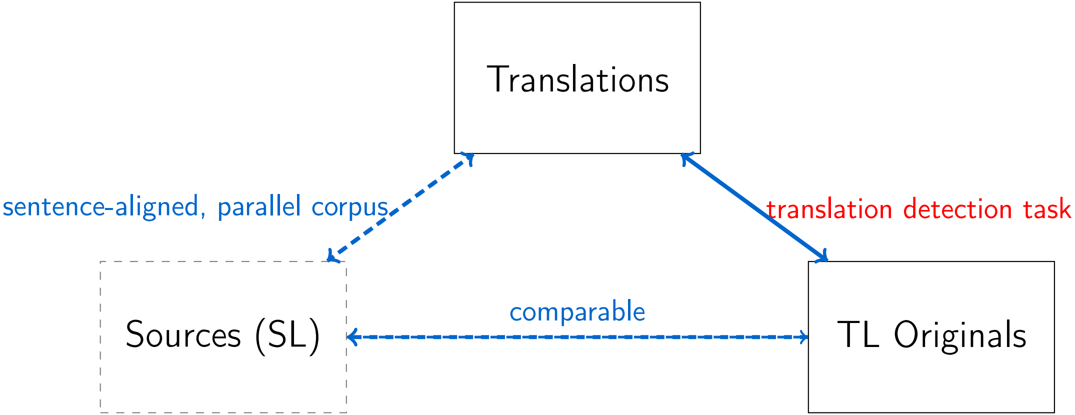
Mediated language –

- is language produced as a result of translation or (simultaneous) interpreting,
- differs in systematic ways from texts originally authored in the target language (TL),
- can be considered a TL subsystem (Chesterman, 2017).

Translationese –

- is systemic deviations of mediated language from the TL norm,
- reflects the specificity of the translation process,
- can be very subtle for an expert human eye (Wein, 2023) but is easily picked up by a machine, esp. at the document level (chunks of 450 or 2000 tokens), with accuracy over 85% on de-lexicalised features.

Typical resources for a translationese study



Methodological paradigms

1. Statistical tests, single-feature classifiers (univariate analysis)
   - that-complementizer (Olohan, 2001)
   - phrasal verbs (Cappelle and Loock, 2017)
   - conjunctive markers, TTR, contracted forms, loan words (Redelinghuys and Kruger, 2015)
2. Text classification on feature vectors (multivariate approach)
   - on hand-crafted theory-based features (Evert and Neumann, 2017),
   - on n-grams (Popescu, 2011; Baroni and Bernardini, 2005; Lapshinova-Koltunski and Zampieri, 2018),
3. Neural feature-learning classifiers to solve
   - translation detection task (Pylypenko et al., 2021)
   - SL detection task (Sominsky and Wintner, 2019; Dutta Chowdhury et al., 2021)

early 1990s corpus analysis → *approx*2006 SVM+features → *approx*2018 end-to-end learning → LM approaches

Suggested translationese trends (Chesterman, 2004)

S-universals induced by source language (SL)

1. interference = ‘shining through’ effect (Teich, 2003)
   translations follow source text (ST) rather than target language (TL) patterns
2. explicitation (Olohan, 2001)
   spelling things out rather than leave them implicit
   - more connectives;
   - more elaboration in brackets;
   - ST non-finite clauses > TT finite clauses (Bisiada, 2014);
   - ST pronouns and ellipsis > TT full NPs (Zinsmeister et al., 2012)
3. levelling-out (aka Standardization/Convergence)
   translations are more homogeneous and less creative than ST
4. lengthening: *translations are longer than their sources*

Suggested trends, cont.

T-universals induced by the gravitational pull from the TL

1. simplification

   less varied vocabulary (Bizzoni et al., 2018), higher readability scores, less figurative language

2. (over)normalization

   tendency to exaggerate properties of the TL; e.g. lexical "teddy-bears"

3. unique items hypothesis

   TL specific items are under-represented

‘Shining-through’ examples

Exploiting intersecting patterns (English -> Russian)

source past rhetoric

translation ritorika proshlogo [rhetoric of the past]

expected staraya ritorika [old/past rhetoric]

Word order interference (English –> German)

source As legislators, however, we must always ... base legislation on sound science.

interpreted Wir als Gesetzgeber müssen uns stützen auf vernünftige wissenschaftliche Ergebnisse.

expected Wir als Gesetzgeber müssen jedoch stets bestrebt sein, ... der Gesetzgebung eine solide wissenschaftliche Basis zu geben.

Surprisals from n-gram LMs as classifier features

(Rubino et al., 2016)

Aim: Classify originals vs translations and professionals vs students (sentence-level) to explore features effectiveness

Data: train: TL-original debates, fiction and politics (9-155K sentences), test: (out-of-domain) professional, student and original (6K sentences) (EN>DE)

Method: 225-dimensional dense vector of $-logP(x)$-based features from [1;5]-gram LMs on word, PoS, flattened syntactic trees

Result: Surprisal is 2nd best, lexical and syntactic text representations perform better than PoS.
Original vs professional accuracy 69.2% (unigram frequency baseline: 78%)

Entropies from frequencies as (sub-)language property

(Nikolaev et al., 2020)

Are translations easier or more difficult to read than originally-authored texts in the TL?

Aim: Explore morphosyntactic predictability of translations wrt distance between SL and TL.

Data: EN>8 languages (4 Indo-European + 4 non-Indo-European)

Method: Produce frequency lists for each *UD deprel* and *PoS*, calculate $P(x)$ as relative frequency (n_x/N)
– with bootstrapping to obtain boxplottable distributions of differences in entropies between TL originals and translations

Result: structurally-similar sources result in decreased entropy of translated language, and structurally-divergent SLs generate more entropic translations.

Entropies from frequencies as classifier features

Entropy of grammar rules and connectives for SVM: (Hu and Kübler, 2021)

Aim: Search for translationese universals and SL interference

Data: 7 source languages > Chinese vs. original Chinese (2,000 tokens chunks)

Extraction: For each chunk, get frequency lists of (i) connectives and (ii) rules headed by each a phrase (NP, VP, etc.) from syntactic annotation.
Calculate $E(XP) = -\sum_{r \in rulesheadedbyXP} p(r) \log p(r)$, where $p(r)$ is the relative frequency (n_x/N) of an item from the list.

Result: – Grammar rules entropies: accuracy 89.27%, markers - 87.87% (but simpler features return even better results).
– Translations overuse cohesive markers and have more complex grammar.
– A strong influence of SL-TL distance!

B7: Translation as rational communication

B7 Goal

a unifying explanation of the observed effects on the grounds of rational communication:

- Can departures from TL norms in mediated language be viewed as efficient information management?

  = Do translation solutions achieve the expected result in the most economical way, thus, striking the balance between effectiveness and efficiency?

Expected result = effectiveness = aspects of translation quality: adequacy, accuracy and fluency.

Entropy and surprisal from translation solutions space

(Martínez and Teich, 2017) See also (Wei et al., 2022)
Translation space is all the available solutions for a SL item from multi-parallel data.
Entropy is the degree of uncertainty of the solution (non-routine translation).
Surprisal is interpreted as acceptability of learner translation against professionals as reference: the same or lower surprisal than in professional space is expected.

Data: Multi-parallel learner corpus with annotated translation difficulties + Professional translations (same register)

Method: Extract translation difficulties and make a frequency list of solutions. Get entropy and surprisal for each solution using $P = n_x/N$ from the list.

main findings

- Learners generate more entropic spaces, their solutions are less routine.
- Professionals show more confidence in solving problems annotated as difficulties: their solutions have lower surprisal and translation spaces have clear preferred solutions.

(Relative) perplexity and (sub-)language properties

Method: Train LSTM on originals, represented as PoS sequences (3K sents/250K tok)
Compare $PP = 2^{H(q,p)}$ on *unseen* tests in SL, TL and mediated language

(Bizzoni et al., 2020)
* train: 6 LMs on spoken/written for original EN, DE, mediated DE,
* test: human and MT for written/spoken.

main findings

- more of structural shining-through in MT than in HT,
- strong normalisation in written EN>DE,
- The specificity of interpreting depends on the cognitive overload of human translators, not the difficulty of speech.

(Bizzoni and Lapshinova-Koltunski, 2021)
* train: 7 registers EN<>DE,
* test: professional and student translations

main findings

- Professionals are more perplexing because they are more register-aware, less homogeneous.

Surprisal and translation solutions wrt connectives

(Lapshinova-Koltunski et al., 2022)
* 4-gram models trained on originals+mediated for each EN and DE mode separately,
* surprisal for pre-selected connectives manually annotated in parallel written/spoken HT for translation strategy (equivalence, implicitation and explicitation)

Expectations:

- Easier equivalence or implicitation strategies are used for ambiguous SL items and more often in interpreting as a coping mechanism.
- Explicating solutions benefit the recipient (and should return lower surprisal) but require more effort, hence not used for weaker relations (ambiguous) markers.

main findings

- More explicitation in translation and more implicitation (∅-connectives) or more general connectives (*but vs. however*) in interpreting.
- Explicitation returns higher surprisal (=more effort for the listener!), compared to equivalent renditions for contrast but not contingency (*because, well*) markers.

Intermediary summary

Typology of research designs using surprisal, entropy, perplexity in TS

Derivation: indices from LM or from frequency lists

Indexed item: text category (subcorpus), document, sentence, token

Usage: direct comparison of indices vs. features for a classifier

Goals and hypotheses

Do the conditions of the mediation influence the information density (ID) of the message?

(using average segment surprisal (AvS) from ngram LM as a measure)

Mediation conditions

- written, spoken
- spoken: read out vs impromptu sources
- spoken: delivery of source (words/min)

Comparison methods

- regression,
- correlation analyses,
- statistical tests

Expectations:

- Positive association between sources and targets
- Lower AvS in spoken than in translation (for the same input)
- More information out in intuitively easier conditions (impromptu, low speaker's speech rate)
- Lower AvS in mediated than in non-mediated language

(Kunilovskaya et al., 2023)

Experimental Setup

Data: 12 K segments, ca. 250 K tokens in each language (Europarl EN<>DE, balanced modes and directions, lemmatised, ST by native speakers)

Modelling decisions

- Each LM uses all available data
- Point-wise probabilities are based on the counts from the entire corpus, excluding the current document,
- OOV: backoff to lower-order n-grams,
- *Hapax legomena* $\rightarrow$ *UKN*,
- n-grams respect sentence boundaries.

4-gram statistical language models for English and German

↓

Average surprisal (AvS) across segment tokens

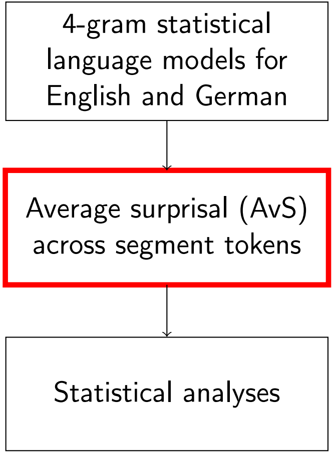
↓

Statistical analyses

Experimental Setup

Surprisal as a measure of information density.

$$S(w_i) = -log2(P(w_i|w_{i-3}, w_{i-2}, w_{i-1}))$$



Experimental Setup

Hypotheses Estimation

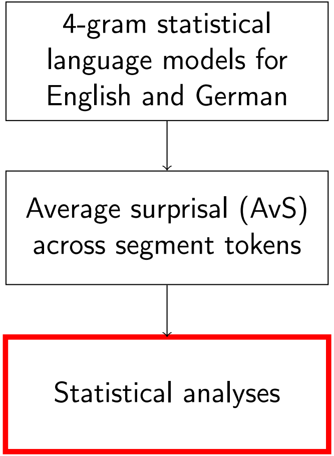
cross-lingual analyses of AvS for aligned segs

- linear regression,
- correlation analyses (Spearman r)

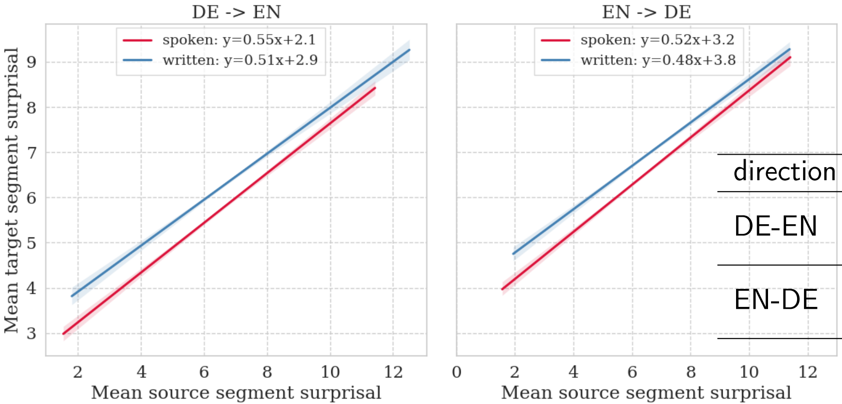
monolingual comparisons –

- statistical significance tests

NB! avoid direct comparisons of AvS values from language-specific models!



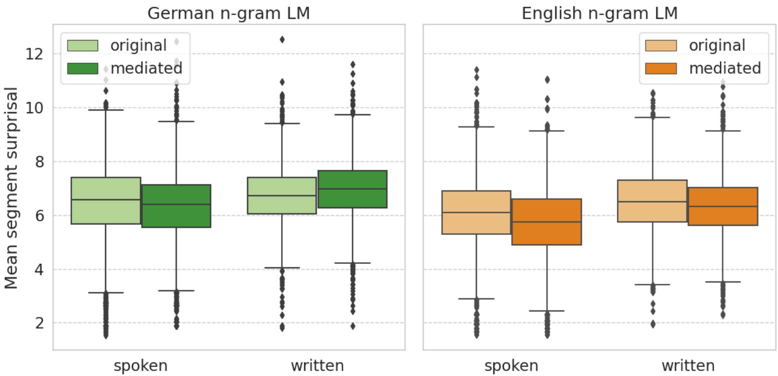
Association sources – targets



direction	mode	r
DE-EN	written	0.47
	spoken	0.48
EN-DE	written	0.51
	spoken	0.44

There is a reasonably strong association between sources and targets.
Given the same information input, translators produce more informative output.
Slope: +0.5 bit in output per 1 bit of input. → Mediated language is denser than sources!

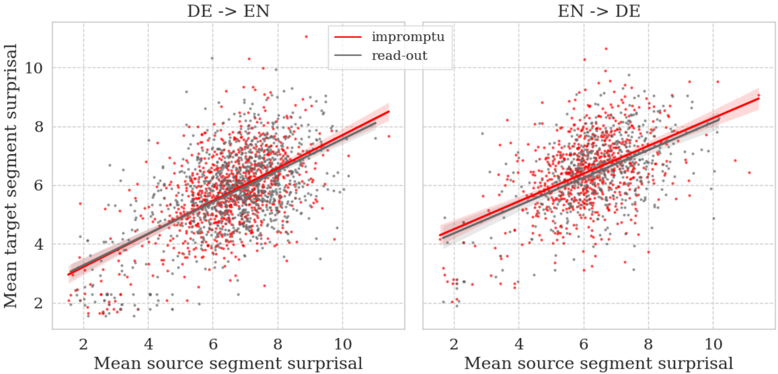
Are mediated texts simpler than comparable TL originals?



Mediated language has lower levels of surprisal than TL originals (except for written into German).

All results are statistically significant.

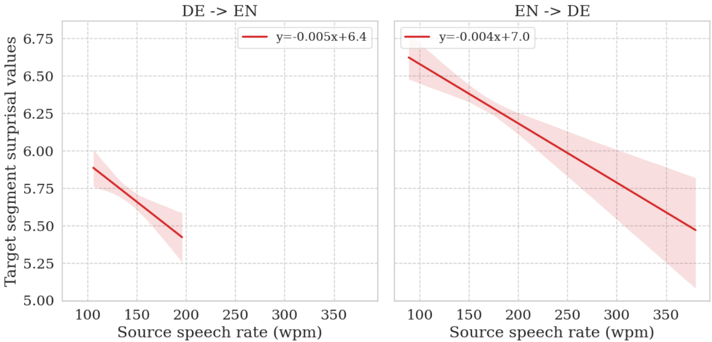
Impact of Challenging Conditions: impromptu vs. read-out



No evidence that read-out sources lead to higher-surprisal targets.

Impact of Challenging Conditions: source speech rate

NB! word-per-minute speed of source speeches as the explanatory x-variable.



Spearman r -0.06 and -0.09 (low, but significant)
$\rightarrow$ the higher the source speed, the lower the informativity of the target.

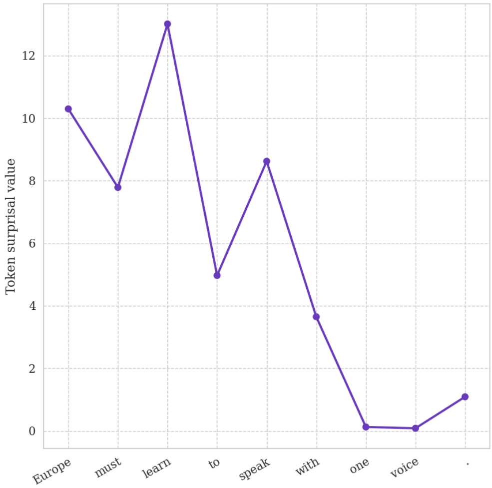
Examples: low-surprisal spoken targets – simplification

(1) **source:** Europa muss lernen, mit einer Stimme zu sprechen und dann auch mit einer Position zu handeln.

translation: Europe must learn to speak with one voice and to take united action.

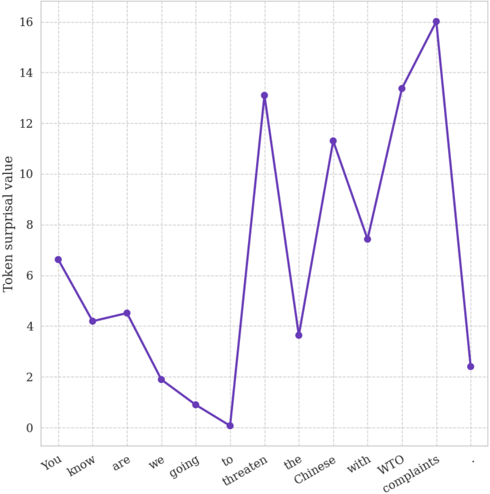
interpreting: Europe must learn to speak with one voice. (AvS = 5.52)

NB! The segments omitted and added in interpreting had lower AvS than the average across all segments in both language directions.



Discussion: high-surprisal spoken targets – interference

(2) **source:** Wollen wir den Chinesen mit WTO Klagen drohen.
translation: Do we want to threaten the Chinese with World Trade Organisation (WTO) sanctions?
interpreting: You know are we going to threaten the Chinese with WTO complaints (AvS = 6.57).



To sum up

Findings

- the informativity of the target is strongly and positively correlated with the informativity of the source, regardless direction and mode,
- the information output in interpreting is lower than in translation given the same input,
- mediated language is simpler than comparable non-mediated speeches in the TL (except translations English-to-German), but higher than in aligned sources,
- the impact of the challenging conditions was barely captured.

Quest for the best modelling approach

data: the same Europarl (big) written and (small) spoken corpora,

preprocessing options: raw tokens, Stanza tokens, Stanza lemmas,

setup: train on originals (3M tokens) separately for each language, test on mediated sets and held-out originals (250K tokens)

models: KenLM 4gram, LSTM 4grram, (pretrained) GPT2 for EN (Schweter, 2020) and DE(v1.0) (Radford et al., 2019), and multilingual GPT2

Internal evaluation: Perplexity

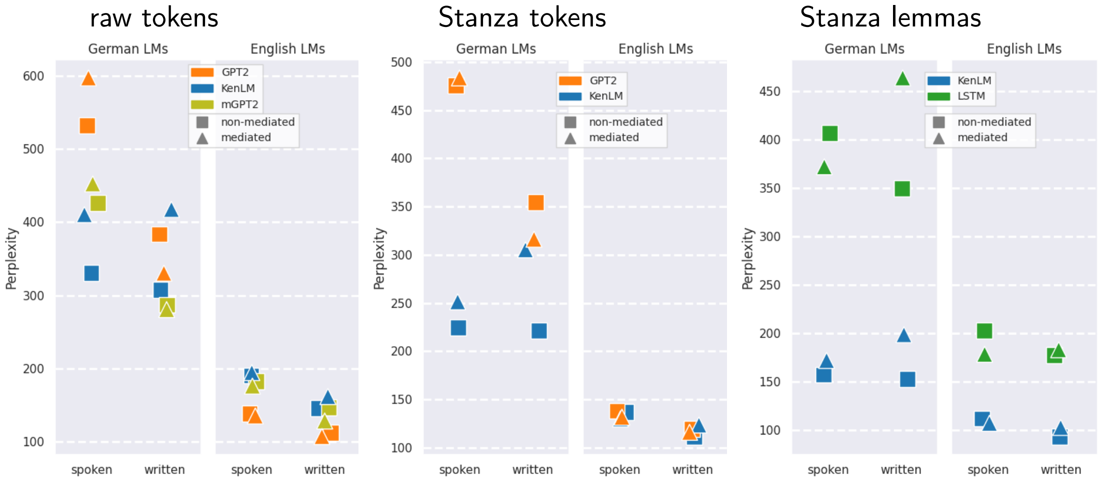
$perplexity(W) = 2^{H(W)}$
where
$H(W) = \frac{1}{N} \sum S(w_i)$

(Jurafsky and Martin, 2023)

External evaluation: Translation detection

- SVM with linear kernel
- Baseline: 4 known translationese indicators
- Surprisal-based features: mean, min and max, standard deviation
- Granularity: seg and doc levels

Internal evaluation: Preprocessing Impact



Voc DE: 116K, EN: 56K

Voc DE: 97K, EN: 40K

Voc DE: 74K, EN: 30K

External evaluation: binary SVM classification

Best-performing approaches vs baseline (min seg. length: 5)

Segment level

		baseF1	F1	best setup
de	sp	55.24	54.19	KenLM(tok)
	wr	53.41	56.72	KenLM(tok)
en	sp	52.17	53.63	LSTM(lem)
	wr	50.76	50.7	GPT2(detok)

Document level

		baseF1	F1	best setup
de	sp	75.43	67.91	KenLM(tok)
	wr	66.32	69.27	KenLM(tok)
en	sp	67.3	62.85	LSTM(lem)
	wr	63.07	55.03	KenLM(tok)

Translated German is more detectable than translated English.
For traditional features spoken is more detectable, for surprisal-based features - no consistency.

Summary

1. Surprisal/entropy features **from word-based LM are inferior** to other features in translation detection classification tasks.
2. More **coarse-grained** setups based on entropy/perplexity measures for entire languages/documents are more promising.
3. Surprisal-supported linguistic analysis of fine-grained translationese phenomena requires **annotation of the focused translation solutions** or multiple translations.
4. Non-LM approaches (entropy from frequency lists), including defining the space of translation solutions, are under-researched.
5. There is a **lack of psychometric experiments** to shed light on what info-theoretical metrics actually capture in the context of cross-lingual communication, where production and comprehension costs need to be distinguished.

Instead of future work

07 June 2024

Multilingual Modelling Workshop

Current B7 research question:

Is non-native language use by native-speaking interpreters a reflection of **working memory management** accommodating for **the source language content**?

- use memory-surprisal estimates from multilingual Transformers to explain trends in translational behaviour

All SFB projects working with, or interested in, multilingual models are welcome to present their work and participate!

Thank you!

Application of Information Theory
in Translation(ese) Studies

Questions?

Maria Kunilovskaya
maria.kunilovskaya@uni-saarland.de

References I

Baroni, M. and Bernardini, S. (2005). A new approach to the study of translationese: Machine-learning the difference between original and translated text. Literary and Linguistic Computing, 21(3):259–274.

Bisiada, M. (2014). The impact of editorial guidelines on sentence splitting in german business article translations. Applied Linguistics, 37(3):354–376.

Bizzoni, Y., Juzuk, T., Espã na Bonet, C., Dutta Chowdhury, K., van Genabith, J., and Teich, E. (2020). How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In Proceedings ofthe 17th International Conference on Spoken Language Translation (IWSLT), pages 280–290. Association for Computational Linguistics.

References II

Bizzoni, Y. and Lapshinova-Koltunski, E. (2021). Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students? In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 53–63, Reykjavik. Linköping University Electronic Press.

Bizzoni, Y., Przybyl, H., and Teich, E. (2018). Cutting semantic corners? Patterns of lexical simplification in interpreting vs. translation. UCCT 2021: Book of Abstracts, pages 20–23.

Cappelle, B. and Loock, R. (2017). Typological differences shining through : The case of phrasal verbs in translated English. In De Sutter, G., Lefer, M.-A., and Delaere, I., editors, Empirical Translation Studies. New Theoretical and Methodological Traditions, pages 235–264. Mouton de Gruyter.

Chesterman, A. (2004). Hypotheses about translation universals. Claims, Changes and Challenges in Translation Studies, pages 1–14.

References III

Chesterman, A. (2017). Reflections on Translation Theory: Selected papers 1993 – 2014.

Dutta Chowdhury, K., España-Bonet, C., and van Genabith, J. (2021). Tracing source language interference in translation with graph-isomorphism measures. In Mitkov, R. and Angelova, G., editors, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 375–385, Held Online. INCOMA Ltd.

Evert, S. and Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. Empirical Translation Studies: New Methodological and Theoretical Traditions, 300:47–80.

Hu, H. and Kübler, S. (2021). Investigating Translated Chinese and Its Variants Using Machine Learning. Natural Language Engineering, 27(3):339–372.

Jurafsky, D. and Martin, J. H. (2023). Speech & language processing.

References IV

Kunilovskaya, M., Przybyl, H., Teich, E., and Lapshinova-Koltunski, E. (2023). Simultaneous Interpreting as a Noisy Channel: How Much Information Gets Through. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, pages 608–618. INCOMA Ltd.

Lapshinova-Koltunski, E., Pollkläsener, C., and Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. The Prague Bulletin of Mathematical Linguistics, 119:5–22.

Lapshinova-Koltunski, E. and Zampieri, M. (2018). Linguistic features of genre and method variation in translation: A computational perspective. The grammar of genres and styles: from discrete to non-discrete units, 320:92–117.

References V

Martínez, J. M. M. and Teich, E. (2017). Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Cercel, L., Agnetta, M., and Lozano, M. T. A., editors, Kreativität und Hermeneutik in der Translation, pages 403–427. Narr Francke Attempto Verlag.

Nikolaev, D., Karidi, T., Kenneth, N., Mitnik, V., Saeboe, L., and Abend, O. (2020). Morphosyntactic predictability of translationese. Linguistics Vanguard, 6(1):1–12.

Olohan, M. (2001). Spelling out the optionals in translation : a corpus study. UCREL Technical Papers, (13):423–432.

Popescu, M. (2011). Studying translationese at the character level. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2011), pages 634–639.

References VI

Pylypenko, D., Amponsah-Kaakyire, K., Dutta Chowdhury, K., van Genabith, J., and España-Bonet, C. (2021). Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8596–8611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Redelinghuys, K. and Kruger, H. (2015). International Journal of Corpus Linguistics.

Rubino, R., Lapshinova-Koltunski, E., and van Genabith, J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. In Proceedings of NAACL HT 2006, pages 960–970, San Diego, California.

References VII

Schweter, S. (2020). German gpt-2 model.

Sominsky, I. and Wintner, S. (2019). Automatic Detection of Translation Direction. In Mitkov, R. and Angelova, G., editors, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Teich, E. (2003). Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. Walter de Gruyter, Berlin/Boston.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

References VIII

Wein, S. (2023). Human raters cannot distinguish English translations from original English texts. In Bouamor, H., Pino, J., and Bali, K., editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12266–12272, Singapore. Association for Computational Linguistics.

Zinsmeister, H., Dipper, S., and Seiss, M. (2012). Abstract pronominal anaphors and label nouns in german and english: Selected case studies and quantitative investigations. Translation: Computation, Corpora, Cognition, 2(1).