Beyond Rankings: Insights from ALTAE's PTT-MULTITAN Submission to the WMT 2026 Shared Task on Automated Translation Quality Evaluation Systems

Lichao Zhu **Maria Zimina-Poirot** **Nicolas Froeliger**

ALTAE, Université Paris Cité
{lichao.zhu, maria.zimina-poirot, nicolas.froeliger}@u-pariscite.fr

Abstract

We describe PTT-MULTITAN, our submission to the WMT26 MT Evaluation shared task, Task 3 (error-free segment detection). The task requires a binary decision – whether a machine translation (MT) contains any error – without access to reference translations, and is scored using the *Matthews Correlation Coefficient* (MCC). Our system combines a supervised classifier trained on WMT25 human error-span annotations with three families of reference-free indicators: structural and phraseological cues, corpus-linguistic association measures computed over an LLM-as-judge model, and format/locale conformity checks. Our improved submission raises MCC from 0.159 to 0.168. We submitted six language pairs from the official WMT26 test set: English into Belarusian, German, Russian, Ukrainian, Simplified Chinese, and Traditional Chinese. Overall, TRIVE emerges as the strongest general-purpose system, with the highest full-coverage score and strong performance across European languages. *Correspondence Analysis* (CA) offers complementary insights into lexical and domain differences, while future evaluation should involve translators and translation scholars.

1 Introduction

1.1 Background

Reference-based metrics such as BLEU (Papineni et al., 2002) require a reference translation, whereas neural metrics such as COMET (Rei et al., 2020), their reference-free variants such as COMETKiwi (Rei et al., 2022), and quality estimation methods function without this requirement, but all predict a graded score. However, all of these approaches ultimately produce a *graded* quality score. In contrast, our task requires a *binary* judgment: TRUE or FALSE. This difference is consequential and directly shapes our strategy (see §4.1). At the same time, machine translations are qualitatively evaluated by human annotation protocols that differ. Direct assessment gives one number per segment, but does not indicate what is wrong. Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) marks every error span with an error category and severity. ESA (Kocmi et al., 2024) offers a middle ground: annotators first mark problematic spans with a binary severity label (minor/major) and then assign an overall segment score.

1.2 Motivation

However, the very question "What counts as error-free, and what does not?" remains extremely complex. Translation studies have argued over it for decades and proposed systematic criteria for what qualifies as "error-free" or "publishable". Descriptive approaches, for instance, locate judgement in norms rather than in the text itself: a shift toward target-language acceptability that one evaluator marks as a deviation, another may treat as expected behaviour (Toury, 1995). Likewise, whether a translation is assessed as overt or covert changes what counts as a fault (House, 2015). To our knowledge, no substantial Natural Language Processing (NLP) study addresses these finer-grained aspects of the problem. As this is a paramount subject, although slightly outside the scope of this exercise, we feel we should go back to it in the discussion.

2 Related Work

A related line of research concerns phraseological features identified through corpus linguistics. Fluent production largely reuses recurrent word combinations (Sinclair, 1991), and mutual information (Church and Hanks, 1990) was the first widely adopted measure of this recurrence. The *logDice* (Rychlý, 2008) was developed for lexicographic

Proceedings of the Eleventh Conference on Machine Translation, pages 1763–1771
October 28-29, 2026 ©2026 Association for Computational Linguistics

applications, designed to be independent of corpus size. A related strand asks not how strong an association is but which words are *collocational*, measuring this through the entropy of a words collocate distribution (Kilgarriff et al., 2004). We adapt this idea to assign weights to unattested *n*-grams based on the fixity of the slot they fill. To our knowledge, such measures have not previously been used as reference-free MT evaluation features before.

3 Task and Evaluation Settings

Task 3 is scored using the *Matthews Correlation Coefficient*, which measures the correlation between predictions and ground truth, unlike accuracy and F_1. Rather than treat this as an implementation detail, we took it as the starting point of the process, since two of its properties are related to what a system should aim to achieve. First, it cannot be satisfied by hedging.

4 Our Approach

Any constant prediction leaves the coefficient undefined, and it is conventionally assigned zero, which is also the value a random predictor attains in expectation. The second is that, being computed on hard decisions rather than on a ranking, it depends directly on where the threshold falls. If the true positive rate is p and a system predicts positive at a rate $r > p$, then even a perfect ranking is capped at $\sqrt{p(1-r)/r(1-p)}$. With $p = 0.25$, predicting at $r = 0.50$ caps the score at 0.58 regardless of how well the hypotheses are expressed. Calibration is therefore essential here. Because a segment counts as error-free only if nothing is wrong with it, the two directions of evidence are not symmetric: a single fault is enough to place a segment in the negative class, whereas placing it in the positive class would require establishing the absence of *every* fault, which no reference-free signal can do.

4.1 Combining MCC metrics with heuristics

The first way to identify defects relies on **heuristics** that a human annotator can use to mark errors without hesitation, such as translation in the wrong language, a verbatim copy of the source, or an empty output. The second addresses issues of meaning and linguistic norms. To this end, we added a prompted LLM judge to the evaluation framework (see Figure 2), iteratively refining its prompt by inspecting cases where its verdict diverged from that of the classifier. The iterative process was based on the assumption that disagreements between the two judgments are precisely where the two sources of evidence can provide the most information about each other.

4.2 The decision procedure

The asymmetry of the strategy outlined in §4.1 translates into a two-stage decision. As one fault is sufficient to disqualify a segment, while some evidence should be allowed to decide independently, we distinguish between a *veto* and a *graded score*.

4.2.1 The veto

A hypothesis is disqualified, regardless of its score, if it has normative issues: its detected language is not the target language, it is empty, its most frequent token exceeds 50% of its tokens, or, for Chinese, the majority of its script-discriminating characters belong to the wrong script, etc.

4.2.2 Score and threshold

All remaining features are aggregated into a continuous score, thresholded separately per language pair. We use two instantiations depending on whether labelled training data covers the language pair.

Supervised: Where WMT25 ESA annotations cover the pair ℓ, a calibrated classifier estimates the probability that a segment is error-free, and the threshold τ_ℓ is the value maximising MCC on out-of-fold probabilities, with a global fallback where a pair has too few examples (< 200) or only one class is present.

Unsupervised: Otherwise the ranking score is `COMETKiwi`'s estimate (Rei et al., 2022) where available, and a composite of penalty terms,

$$\text{score} = -\sum_k w_k \, P_k \tag{1}$$

where P_k denotes the structural and phraseological penalties, and w_k are fixed weights, cut so that a target proportion is predicted error-free.

Both instantiations prioritise the veto, so the veto can only remove positives, never add them.

4.2.3 Penalties

We have come across, between two submissions, the very weak performance on Chinese and Russian, which prompted us to adapt our strategy

Pair	Prior	v1	v2	v3	v4
Per-pair MCC					
bel_Cyrl	0.189	**0.262**	0.173	0.132	0.226
deu_Latn	0.255	**0.247**	0.182	0.100	0.126
rus_Cyrl	0.274	0.138	0.206	0.158	**0.208**
ukr_Cyrl	0.241	**0.252**	0.236	0.184	0.213
zho_Hans	0.308	0.055	0.120	**0.183**	−0.012
zho_Hant	0.265	−0.001	0.092	**0.145**	0.046
Macro MCC	–	0.159	**0.168**	0.150	0.134
F_1	–	0.300	**0.439**	0.415	0.398

Table 1: All scored versions.[1] Per-pair figures are MCC, recomputed from the released confusion matrices; F_1 is measured against the automatic pseudo-reference (column **Prior**).

by introducing ratio, structural and phrasological penalties. The features measure conformity rather than correctness. An accurate but unusual translation is thus penalised, whereas an error shared by every system in the pool is considered as "usual".

5 Training Data

We used a LIGHTGBM classifier (`n_estimators=400`, `num_leaves=31`, `class_weight="balanced"`). It was recalibrated with an isotonic regression by a 3-fold cross-validation, and was fitted on the WMT25 General MT human evaluation release, which is one of the rare resources whose labels can be reconstructed for Task 3 training. Each record covers one source segment and maps every system to a list of *Error Span Annotations*, where each annotation carries a segment-level score, a list of error spans with character offsets and a `minor/major` severity (Kocmi et al., 2024). For the training data, where a pair carries several annotations, we adopt a *conservative* aggregation: the pair counts as error-free only if every annotation says so, as we consider the identification of a span by a single annotator sufficient to classify it as a defect.

Of the 16 pairs covered by the release, only three directly overlap with ours (`en-ru`, `en-uk`, `en-zh_CN`): Belarusian and German from English are absent, and Traditional Chinese is represented only through its simplified counterpart. Thus, supervision is in-domain for half of our language pairs and transferred to the remaining pairs.

[1] Only v2 and v3 were submitted.

6 Results

6.1 Scored submissions

We improved our system across submissions, with the MCC increasing from 0.159 to 0.168 (see Table 1). The improvement from v1 to v2 is not a uniform gain. The structural and phraseological penalties introduced in v2 raised the score of three pairs: `ukr_Cyrl` ($+0.068$), `zho_Hans` ($+0.065$) and `zho_Hant` ($+0.093$), while lowering it on `bel_Cyrl` (-0.089), `deu_Latn` (-0.065) and `ukr_Cyrl` (-0.016). Therefore, the penalties act like a recalibration rather than an additional filter. The two Chinese pairs behave differently from European ones, and we attribute it to the interaction between our features and character-based tokenisation. As phraseological statistics are computed over bigrams for Chinese too, the association of two Chinese characters is measured far more densely than the bigrams used in European languages, since the Chinese is space-free.

The whole system predicts positive at a rate which is above the true one – 57% to 69% of segments, compared with the true rates (priors) which range from 19% to 31%. Thresholds are fitted on WMT25 ESA, which is one of the main parameters we can act on. The gap between the calibration and evaluation distributions, rather than the ranking features themselves, seems to be the source of loss.

6.2 System performance evaluation and ranking

As we have explained in the introductory remarks, the benchmark we explored covers six English-to-target-language pairs: Belarusian, German, Russian, Ukrainian, Simplified Chinese, and Traditional Chinese.

In this study, translation systems with 2,451 translated segments can be compared directly, whereas systems tested on smaller subsets require separate interpretation. This distinction is essential because language coverage affects the results.

According to our evaluation, **TRIVE** provides the strongest overall result among full-coverage systems (see Table 2 and Appendix A). Its 82.6% score exceeds **Wayfinder** and **Lumen** by 0.6 percentage points and GPT 5.5 by 0.7 points. The next tier comprises **MakotoAI** at 81.3%, Gemini 3.1 Pro at 80.0%, and Qingqiu-MT-9B at 79.5%. These results suggest that several translation systems are quite close in performance, although their

language-specific profiles differ.

Based on our results, the leading systems perform particularly well on the four European language pairs. **Wayfinder** obtains the best score for Belarusian (94.1%) and the best score for Russian (91.1%). **MakotoAI** and Lumen share the lead for German at 91.1%, while **TRIVE** leads for Ukrainian at 91.6%. **TRIVE** remains consistently strong across this group, at 88.7% for Belarusian, 88.9% for German, 85.2% for Russian and 91.6% for Ukrainian. This European-language strength seems a major driver of its first-place ranking.

Our results for Chinese are quite different. **Gemma 4 - 31B** leads for Simplified Chinese at 73.0%, slightly ahead of **Cohere CAT+** at 72.6%, and also ranks first for Traditional Chinese at 80.3%. **TRIVE** is second for Traditional Chinese at 79.4% but reaches only 64.2% on Simplified Chinese. Similarly, **Wayfinder** falls from 94.1% on Belarusian and 91.1% on Russian to 59.7% on Simplified Chinese. We believe that these differences show that strong general performance does not necessarily imply uniform multilingual robustness.

The results also reveal different system profiles. **TRIVE** offers the best combination of overall performance and coverage (see Table 2 and Figure 1). **Wayfinder** and **Lumen** are exceptionally strong on European language pairs but are less competitive on Chinese. **GPT 5.5** follows a similar pattern, combining high scores for European languages with 63.8% on Simplified Chinese and 70.8% on Traditional Chinese.

By contrast, **Gemma 4 - 31B** has a lower overall score of 75.6% but a relatively balanced profile and the strongest results on both Chinese languages. According to these results, the best system may depend on whether the objective is maximum aggregate performance or more even performance across language families.

Some partial-coverage systems also have high overall scores, but these figures should not be interpreted within the full ranking results. For example, **VoxNexus_V1** reaches 84.0%, higher than every full-coverage system, yet this result is based on only 780 segments, or 31.8% of the complete benchmark, and covers two language pairs. **Yandex** reaches 82.4% on 426 segments, while **Yandex-8B** reaches 80.0% on 780 segments. These results are promising within their tested scope, but they do not reflect superior performance across the entire multilingual evaluation.

System	Score	NSeg	Coverage
TRIVE	82.6%	2451	100%
Wayfinder	82.0%	2451	100%
Lumen	82.0%	2451	100%
GPT 5.5	81.9%	2451	100%
MakotoAI	81.3%	2451	100%
Gemma 4 - 31B	75.6%	2451	100%
VoxNexus_V1	84.0%	780	80%

Table 2: Comparative performance of translation systems based on our evaluation strategy **pred_v2**.

Overall, **TRIVE** emerges as the most convincing general-purpose system in this evaluation as it combines the highest full-coverage score with strong results across the European language pairs with a leading score for Ukrainian. Nevertheless, no single system dominates every language pair. According to our results, **Gemma 4 - 31B** is the strongest option for the Chinese evaluations, while **Wayfinder**, **Lumen** and **MakotoAI** lead individual European pairs. Therefore, the core takeaway is twofold: aggregate rankings are useful to identify general-coverage solutions, yet the language-specific data and scope must continue to be crucial in determining the best system choice.

7 Discussion

Our results, as well as those of other systems in this shared task, rely on a particular linguistic definition of what constitutes an error and on the assumption that the source and target texts should stand in a relation of equivalence. This may be a reasonable and valid assumption in the context of automated language processing. It is not – or, at least, not systematically and no longer – the case in the actual translation profession, nor, for that matter, in translation studies, at least since the much-discussed "cultural turn" of the mid-1980s. In the outside world, quality is more and more defined first and foremost by the degree of adherence to a translation brief (Reiss and Vermeer, 2014), whether the latter is made explicit or not (and it is the translator's task to have it made explicit). Thus, the same source-language text can yield markedly different outputs in the target language, depending on its intended readers, on their backgrounds, and on the very reasons why they are to read the said output. This should, of course, be taken into account in research such as ours, and this is also one of the reasons we have started introducing a

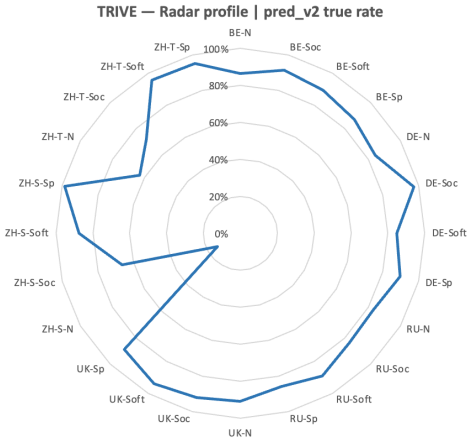

Figure 1: TRIVE's profile in different languages and domains.

prompted LLM judge, which could, with some refinement, go some way towards emulating this translation brief situation.

Within this task-dependent understanding of translation quality, *Correspondence Analysis* (CA) (Lebart et al., 1998) can provide a complementary view of how translation systems handle domain-specific lexical patterns (see Figures 4 and 5). This type of analysis allows comparing the lexical organisation of the source and machine-translated texts across the ***news***, ***social***, ***software***, and ***speech*** domains.

For example, for the English-German language pair, the **pred_v2** results are relatively balanced across the four domains, with the highest score for ***speech*** (75.1%), followed by ***software*** (71.7%) and ***social*** texts (69.7%), while news has the lowest score (62.6%). The overall score is 68.7% across 13,770 segments, suggesting greater difficulty with automatically translated news texts. The CA results provide a complementary view of these differences (see Figures 4 and 5 in the Appendix). While the source corpus shows relatively cohesive domain-specific clusters, the translated corpus exhibits greater lexical dispersion, particularly in the ***speech*** domain. Conversational items such as *hey*, *okay*, *gerade*, and *genau* are distributed across multiple regions of the CA plane, potentially reflecting weaker lexical cohesion and difficulties in preserving conversational features. At the same time, ***news*** and ***social*** texts are mapped closer together in the translated corpus than in the source corpus, suggesting increased lexical overlap between these domains.

8 Conclusion

We presented a reference-free system for detecting translation errors that combines a supervised classifier trained on WMT25 human error-span annotations with three complementary families of indicators: structural and phraseological cues, corpus-linguistic association measures derived from an LLM-as-judge model, and format and locale conformity checks. On the WMT26 test set, this combination yields a modest but measurable improvement over our previous submission, with the MCC increasing from 0.159 to 0.168. We evaluated the system on six language pairs, translating from English into Belarusian, German, Russian, Ukrainian, Simplified Chinese, and Traditional Chinese.

Taken together, these results suggest a gradual rather than revolutionary path toward improving machine translation. As scientific progress often proceeds through incremental steps (Kuhn, 1970, 1994), we hope our work contributes one such step. However, our findings should be interpreted with caution, as they rely on a conception of translation error based on source-target equivalence, raising broader questions about how translation quality should be defined and evaluated.

More specifically, we believe that the interpretation of our evaluation results is limited in the absence of a clearly defined translation brief (Reiss and Vermeer, 2014; Froeliger et al., 2023). Different briefs may prioritise different aspects of performance, such as terminological accuracy, fluency, consistency, speed, or cost. Within the field of translation studies, translation quality is not regarded as an unchanging concept; rather, it is evaluated based on the purpose of the translation, its intended audience, and the role of the target text (Froeliger et al., 2023). Thus, there can be no single machine translation system that is universally the best, regardless of its intended use. Instead, the suitability of a system ultimately depends on the specific translation context and the requirements of the task at hand.

Consequently, benchmark rankings should be interpreted as comparative indicators of performance rather than measures of translation quality. Their practical significance ultimately depends on how well the evaluated systems align with the requirements of the intended translation task. In this respect, CA results can help identify shifts in lexical distribution, domain differentiation, and cor-

	news	social	software	speech	Total
eng_Latn-bel_Cyrl	62.4%	70.8%	72.5%	66.9%	68.4%
eng_Latn-deu_Latn	**62.6%**	**69.7%**	**71.7%**	**75.1%**	**68.7%**
en ag_Latn-rus_Cyrl	67.0%	66.7%	72.3%	70.5%	69.0%
eng_Latn-ukr_Cyrl	63.8%	67.9%	74.2%	66.6%	68.4%
eng_Latn-zho_Hans	18.5%	47.4%	78.6%	89.3%	56.5%
eng_Latn-zho_Hant_TW	28.5%	52.3%	88.8%	89.2%	61.9%

Table 3: Percentage of correct predictions (% VRAI **pred_v2**) by language pair and domain.

pus cohesion, providing useful evidence for assessing translation quality in relation to the translation brief (Zhu et al., 2025). However, CA results should be regarded as exploratory and complementary evidence rather than as standalone measures of translation quality. This type of analysis does not directly assess adequacy, fluency, omissions, or mistranslations. These aspects require human evaluation and additional quality-estimation metrics and should be confirmed through a targeted qualitative error analysis.

In this respect, our contribution should be regarded only as a starting point, and we firmly believe that research such as ours will truly come to fruition once the service-provision dimension is fully taken into account (European Master's in Translation (EMT), 2022), that is, when professional translators and translation studies scholars are involved in the process.[2]

Acknowledgments

This research was funded by the 2024 **MULTITAN-GML** Research Equipment Grant (COPES-2024-12, *Fonds dintervention Recherche, Université Paris Cité*), supporting the **PNS-UP** platform used in this research.

References

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

European Master's in Translation (EMT). 2022. European master's in translation: Competence framework. 2009, revised 2017 and 2022.

Nicolas Froeliger, Christopher Gledhill, and Maria Zimina-Poirot. 2023. Intégrer des plateformes de traduction automatique neuronale dans l'enseignement de la traduction spécialisée. In *Traduction humaine et traitement automatique des langues. Vers un nouveau consensus ?*, number 35 in Studi e ricerche. Edizioni Ca' Foscari – Venice University Press, Venice.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge, London and New York.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105–115, Lorient, France. Université de Bretagne-Sud.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, et al. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Thomas S. Kuhn. 1970. *The Structure of Scientific Revolutions*, 2nd, enlarged edition, volume 2 of *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago. First edition published 1962.

Thomas S. Kuhn. 1994. *The Structure of Scientific Revolutions*, 2nd, enlarged edition, volume 2 of *International Encyclopedia of Unified Science*. University of Chicago Press. 21st printing.

Ludovic Lebart, André Salem, and Lisette Berry. 1998. *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht, Boston, London.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, (12):455–463.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

[2] A lot has been done in this regard since the first **Tralogy** conference, in 2011, but we feel this convergence is not yet complete and should be pursued further.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, et al. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katharina Reiss and Hans J. Vermeer. 2014. *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge, London. Translated by Christiane Nord; original 1984.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN)*, pages 6–9, Brno, Czech Republic. Masaryk University.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Describing English Language. Oxford University Press, Oxford.

Gideon Toury. 1995. *Descriptive Translation Studies — and Beyond*. John Benjamins, Amsterdam and Philadelphia.

Lichao Zhu, Maria Zimina-Poirot, Cristian Valdez, and Stephane Patin. 2025. Fine-tuning NMT models and LLMs for specialised EN-ES translation using aligned corpora, glossaries, and synthetic data: MULTITAN at WMT25 terminology shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1284–1291, Suzhou, China. Association for Computational Linguistics.

A Appendix

```
You are a strict translation quality judge.
Source ({src_lang}): {src}
Translation ({tgt_lang}): {hyp}
Translation domain: {domain}

A translation is ERROR-FREE only if it needs no correction at all: accurate,
fluent, complete, correct style/register/locale.
Answer with exactly one word: PERFECT or FLAWED.
```

Figure 2: Zero-shot prompt for the LLM judge. The `{domain}` field is filled with the domain of the segment (*news*, *social*, *software* or *speech*)

HEATMAP: ALL SYSTEMS × 24 LANGUAGE–DOMAIN COMBINATIONS

System	Belarusian				German				Russian				Ukrainian				Simplified Chinese				Traditional Chinese			
System	News	Social	Software	Speech	News	Social	Software	Speech	News	Social	Software	Speech	News	Social	Software	Speech	News	Social	Software	Speech	News	Social	Software	Speech
Ministral 3 - 14B	4%	0%	38%	47%	1%	1%	48%	68%	1%	0%	55%	63%	1%	0%	51%	56%	4%	1%	56%	79%	10%	0%	83%	80%
CUNI-EdUKate					35%	39%	43%	42%					20%	31%	41%	33%								
Tower 9B	15%	13%	25%	27%	20%	23%	49%	78%	18%	23%	56%	75%	23%	39%	58%	74%	9%	16%	66%	90%	7%	21%	71%	89%
GLM 4 - 9B	23%	39%	25%	41%	31%	45%	44%	64%	44%	43%	45%	40%	26%	51%	46%	42%	15%	30%	65%	85%	4%	33%	83%	76%
Qwen 3.5 - 9b	23%	52%	38%	51%	34%	61%	42%	55%	32%	57%	46%	63%	25%	54%	36%	54%	24%	50%	75%	83%	16%	47%	75%	83%
CUNI-MH-v3					40%	45%	66%	67%					33%	41%	69%	46%								
Qwen 3.6 - 27b	43%	73%	41%	51%	39%	68%	44%	79%	31%	69%	51%	69%	30%	70%	55%	64%	10%	46%	66%	87%	7%	48%	54%	89%
SCIR-TG-MT									37%	51%	54%	57%					33%	41%	70%	88%				
GPT OSS 20B	40%	66%	50%	59%	50%	71%	55%	60%	44%	56%	51%	60%	31%	58%	41%	51%	25%	35%	76%	87%	34%	41%	88%	79%
Mistral Medium 3.5	68%	5%	70%	63%	72%	26%	75%	86%	67%	17%	77%	74%	67%	29%	75%	70%	14%	16%	60%	94%	27%	6%	94%	91%
Tiny Aya Global	31%	38%	68%	23%	56%	60%	75%	67%	52%	45%	78%	57%	62%	44%	79%	52%	31%	30%	75%	74%	24%	14%	96%	83%
ReMova	47%	49%	83%	45%																				
Krosny	46%	51%	69%	67%																				
SRPOL	63%	72%	76%	81%	69%	60%	61%	67%	67%	50%	51%	60%	73%	63%	60%	65%	8%	50%	73%	91%	13%	53%	79%	88%
Gemma 4 - 4B	48%	74%	48%	42%	62%	80%	75%	74%	67%	73%	74%	70%	58%	73%	74%	68%	14%	48%	74%	87%	33%	51%	90%	88%
SalamandraTA	51%	67%	69%	67%	65%	71%	71%	69%	56%	52%	67%	50%	63%	76%	75%	67%	21%	45%	84%	88%	36%	52%	90%	88%
GPT OSS 120B	52%	85%	70%	70%	69%	78%	81%	77%	61%	69%	58%	75%	57%	74%	70%	68%	11%	43%	63%	91%	7%	55%	88%	87%
Human (postediting)					53%	69%	76%	71%																
HW-TSC													86%	83%	91%	72%	14%	42%	81%	74%	6%	60%	96%	95%
GoogleTranslate	81%	88%	90%	79%	51%	41%	75%	69%	82%	73%	70%	75%	81%	87%	82%	57%	8%	42%	76%	94%	51%	66%	81%	89%
Cohere CAT+	51%	70%	74%	55%	71%	75%	84%	71%	75%	71%	75%	55%	76%	72%	86%	68%	57%	67%	86%	79%	50%	62%	94%	74%
Dubformer	74%	91%	90%	60%	83%	93%	88%	64%	71%	85%	88%	49%	75%	88%	89%	53%	19%	69%	87%	74%	6%	66%	94%	77%
Command A+	73%	87%	89%	72%	82%	80%	84%	78%	79%	78%	81%	68%	84%	71%	87%	65%	16%	55%	87%	91%	41%	50%	88%	95%
UvA-MT	81%	88%	84%	86%	69%	88%	62%	78%	82%	77%	75%	81%	78%	87%	89%	73%	14%	61%	86%	92%	11%	73%	94%	94%
PseudoRef	93%	95%	91%	76%	78%	74%	83%	69%	82%	69%	83%	61%	91%	79%	84%	67%	36%	42%	87%	94%	37%	48%	96%	93%
DeepSeek V4 Pro	88%	89%	91%	86%	85%	92%	88%	88%	73%	73%	69%	79%	81%	84%	89%	74%	9%	48%	88%	92%	30%	50%	96%	93%
Gemma 4 - 31B	59%	77%	68%	71%	76%	92%	71%	88%	79%	78%	75%	75%	76%	79%	65%	74%	45%	67%	88%	97%	67%	70%	96%	98%
Qingqiu-MT-9B	79%	91%	87%	81%	72%	92%	84%	88%	84%	88%	86%	85%	81%	83%	93%	84%	15%	66%	87%	96%	64%	66%	94%	95%
Yandex-8B	68%	85%	84%	73%					79%	84%	84%	81%												
Gemini 3.1 Pro	88%	84%	89%	90%	83%	92%	84%	90%	86%	78%	85%	88%	88%	74%	89%	84%	21%	66%	87%	97%	41%	66%	96%	96%
MakotoAI	93%	98%	90%	84%	88%	97%	90%	88%	93%	91%	85%	83%	85%	86%	84%	79%	15%	54%	87%	96%	49%	66%	96%	98%
GPT 5.5	99%	94%	90%	81%	85%	96%	88%	88%	90%	88%	87%	83%	92%	93%	90%	84%	13%	70%	87%	95%	11%	78%	94%	96%
Lumen	98%	94%	92%	84%	90%	94%	91%	90%	96%	90%	87%	85%	90%	90%	89%	81%	11%	55%	87%	96%	24%	76%	96%	95%
Wayfinder	98%	94%	93%	92%	85%	93%	92%	88%	97%	87%	91%	85%	92%	88%	89%	84%	12%	52%	87%	95%	26%	74%	92%	95%
Yandex									82%	80%	84%	82%												
TRIVE	86%	91%	90%	87%	85%	97%	85%	90%	83%	84%	89%	86%	91%	92%	94%	89%	14%	66%	87%	99%	63%	72%	96%	95%
VoxNexus_V1	75%	94%	89%	81%					86%	90%	83%	73%												

Figure 3: System Performance Rate map: from **green** (higher) to **red** (lower)

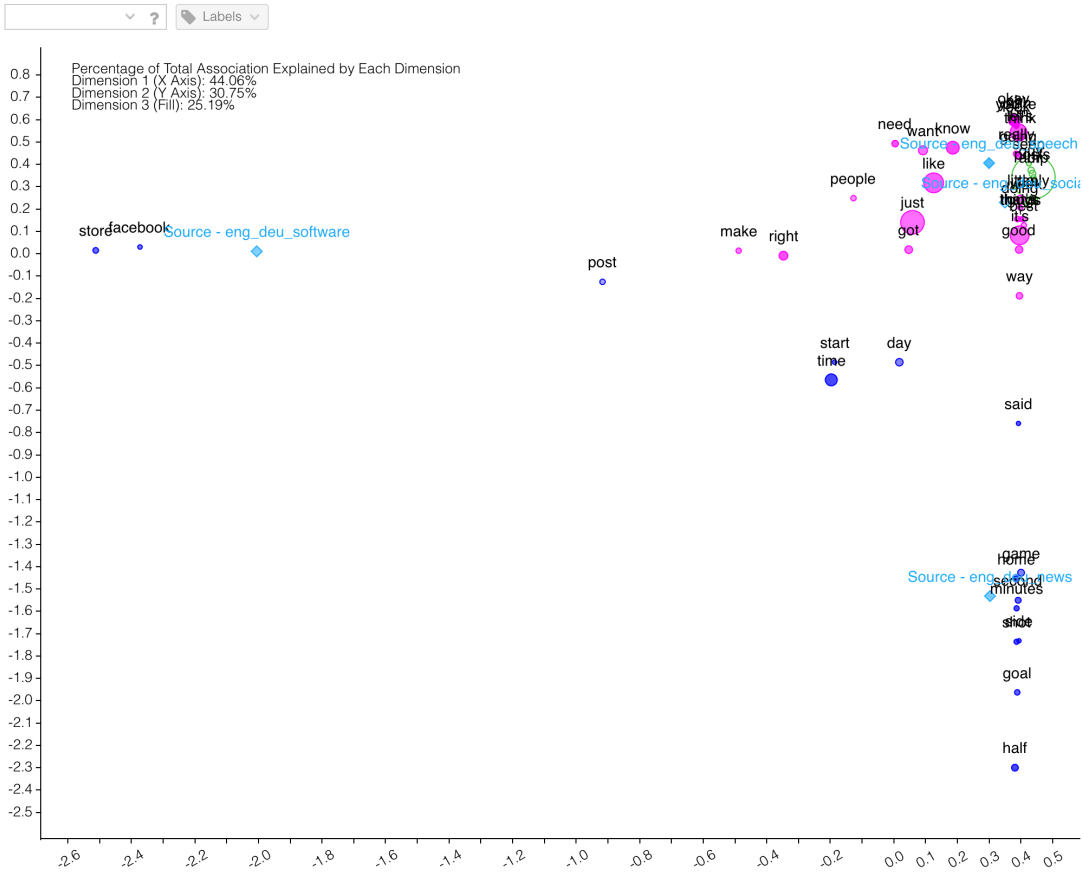

Figure 4: CA: English-German SRC segments across domains

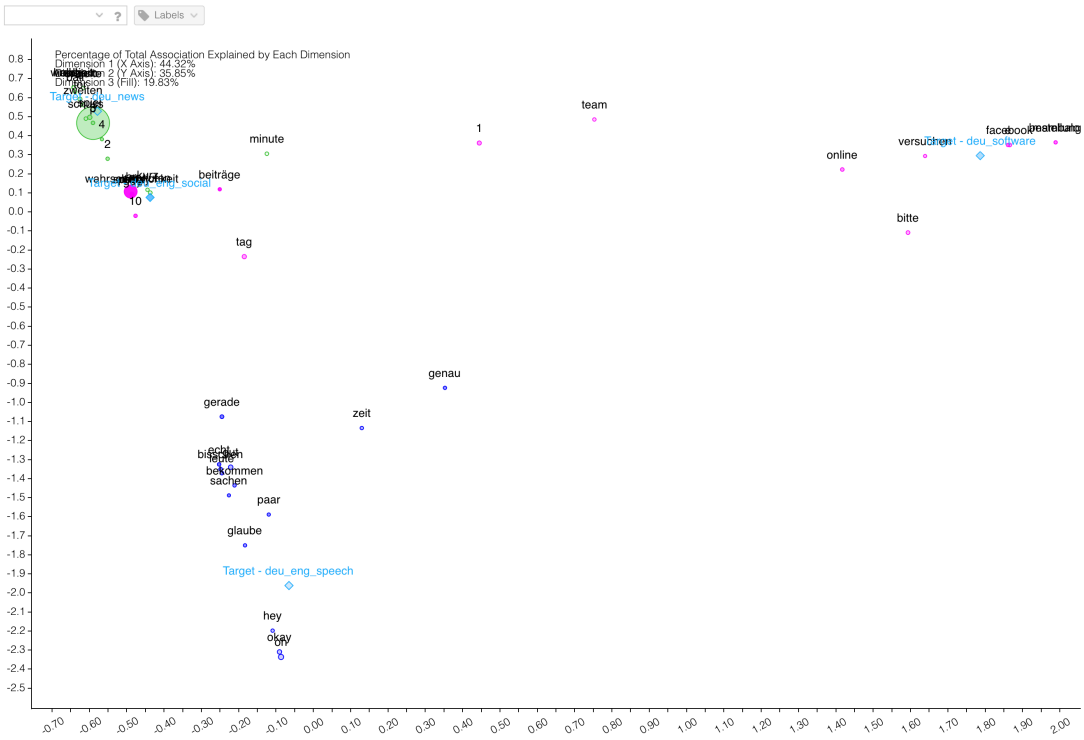

Figure 5: CA: English-German TGT segments across domains