Cultural Grounding in Swedish: Extending an Everyday Knowledge Benchmark for LLMs

Meriem Beloucif* **and Johan Sjons***
Uppsala Universitet
{meriem.beloucif, johan.sjons}@lingfil.uu.se

Abstract

Benchmarks for evaluating Large Language Models (LLMs) on everyday knowledge across cultures and languages are increasingly used to assess cultural competence and contextual understanding. However, many multilingual extensions rely primarily on translated question–answer pairs, limiting their ability to capture locally grounded variation. In this work, we present a Swedish extension of an existing cross-cultural everyday knowledge benchmark, in which questions are translated into Swedish and answers are collected individually from five participants with diverse social and professional backgrounds. This design enables us to capture situated, naturally produced responses from a specific participant group rather than transferred or translated answer templates. We document the translation protocol, participants, and agreement analysis, and examine variation across participants as a signal of culturally contingent knowledge. We evaluate several state-of-the-art multilingual and instruction-tuned LLMs against the aggregated human responses and analyze model performance. Our results reveal that while models often approximate prototypical answers, they struggle with culturally specific nuances and intra-cultural variation. The Swedish extension provides a resource for studying culturally grounded evaluation and highlights the importance of human-generated local answers when benchmarking LLMs across languages.

Keywords: LLM Evaluation, Resource Creation, Cultural Evaluation

1. Introduction

Generative AI has begun to alter how language is produced, interpreted, and evaluated across a wide range of contexts. By early 2026, ChatGPT was estimated to handle approximately 2.5 billion user queries per day.[2] Several researchers (Henrich et al., 2010; Naous et al., 2024; DURMUS et al., 2024) have shown that Large Language Models (LLMs), exhibit characteristics commonly described as **WEIRD: Western, Educated, Industrialised, Rich, and Democratic**.

This is not surprising, given that training data for such models are heavily dominated by English-language content and by textual sources originating from a relatively narrow set of sociocultural contexts. However, even in languages that are morphologically close to English and share socio-cultural and interactional conventions with Anglophone contexts, outputs may still fail to align with local communicative norms and/or expectations, indicating a gap between model outputs and context-specific norms in practice.

To address this gap, we present a Swedish extension of an existing cross-cultural benchmark for evaluating LLMs on everyday knowledge.[3] While Sweden itself falls within the WEIRD category, the mismatch between model outputs and locally grounded practices can still arise, since models surely rely on broadly shared or globalized norms.

The original questions are translated into Swedish by a native speaker, the answers are independently generated by five Swedish participants from diverse social backgrounds, all of whom had Swedish as their first language. It should be noted that all five participants currently live in Stockholm or the Stockholm area, although three of them grew up elsewhere in Sweden. The selection of participants reflects practical constraints in recruiting participants.

The design allows us to capture culturally grounded, naturally produced responses (from our participant group) rather than transferred answer templates. By introducing a Swedish dataset extension with multi-participant, human-generated answers, we contribute a resource for more culturally sensitive benchmarking and provide methodological insights into extending the evaluation of everyday knowledge across multiple languages.

Using this Swedish extension, we evaluate several LLMs to assess how well they approximate Swedish everyday knowledge. Our findings show that none of the models achieves more than 51% matching; most answers are overly generic and appear to be English-based or stereotypical, corroborating our hypothesis about the importance of culturally informed human answers.

* Equal contribution.

[2] https://explodingtopics.com/blog/chatgpt-users

[3] https://github.com/belomeriem/Swedish_BLEND.git

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 25–31 May 11, 2026.

Figure 1: Our process of creating the BLEnD Swedish extension. We have carefully followed the guidelines given by the authors of the original BLEnD paper (Myung et al., 2024).

2. Related Work

2.1. Swedish Datasets

Prior work on Swedish language resources has produced a variety of NLP datasets and benchmarks, though most focus on general linguistic tasks rather than everyday or culturally grounded knowledge. Superlim (Berdicevskis et al., 2023) is a comprehensive Swedish language understanding benchmark modeled after English benchmarks like GLUE, covering multiple NLP tasks to evaluate model proficiency in Swedish contexts.

Other recent efforts include MedQA-SWE (Hertzberg and Lokrantz, 2024), a clinical question-answer dataset designed to assess the domain knowledge of generative models in Swedish medical contexts. Beyond the task of Question Answering (QA), Swedish resources such as linguistic complexity corpora and large web text collections (e.g., SWEb for Scandinavian languages; Norlund et al., 2024) support broader modeling and evaluation work.

Several recent studies have introduced Swedish datasets for complex, semantically motivated tasks, such as semantic relatedness (Ousidhoum et al., 2024), emotion analysis (Muhammad et al., 2025), and Sweden-related facts (Kunz, 2025), and some focus on syntax (e.g., Lundqvist, 2025; Sjons et al., 2026). Our work complements these efforts by extending BLEnD to Swedish, focusing on culturally informed everyday knowledge rather than general linguistic ability or domain-specific expertise.

2.2. LLMs and Cultural Datasets

Large language models (LLMs) acquire extensive parametric knowledge during large-scale pretraining, yet the distribution of that knowledge reflects structural imbalances in the underlying data. Because digital content is unevenly produced across regions and languages, LLMs tend to internalize perspectives that are overrepresented online while underrepresenting culturally specific and locally grounded practices (Bender et al., 2021; DURMUS et al., 2024). These disparities become particularly visible in tasks requiring everyday cultural reasoning, where models may default to globally dominant or Western-centric norms rather than context-sensitive interpretations. A growing body of work has examined cultural knowledge in NLP, often operationalizing culture at the national level and relying primarily on English-language resources (Anacleto et al., 2006).

The current state-of-the-art effort in evaluating cross-cultural everyday knowledge is *The Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages* (BLEnD; Myung et al., 2024), which is carefully human-crafted and covers 13 languages across 16 countries and regions. BLEnD includes underrepresented scenarios and uses aligned question sets to enable direct cross-linguistic comparison. By focusing on everyday knowledge rather than purely encyclopedic content, BLEnD establishes a unified, culturally grounded evaluation framework for multilingual LLMs. Despite this progress, Swedish remains absent from large-scale cross-cultural everyday knowledge benchmarks. Although Sweden has

a strong digital presence, Swedish is typically categorised as a medium-resource language in NLP, since, in contrast to high-resource languages, it has much less parallel data and fewer annotated datasets for complex NLP tasks (Joshi et al., 2020). Moreover, culturally situated aspects of Swedish everyday life – such as institutional norms, seasonal practices and social conventions – cannot be assumed to transfer reliably from other linguistic contexts and closely related high-resource languages.

To address this gap, we introduce a Swedish extension of BLEnD. The original benchmark questions are translated into Swedish using a controlled protocol, while answers are independently collected from five Swedish participants with diverse backgrounds. The design allows us to capture situated, naturally produced responses from our participant group and foregrounds intra-cultural variation as an evaluative dimension rather than inter-participant variation. By extending a state-of-the-art cross-cultural benchmark to a comparatively low-resource language, we enable systematic comparison of LLM performance across languages while improving cultural coverage in multilingual evaluation.

3. Dataset Construction

For the creation of our dataset, we follow the same steps as the BLEnD's authors for data aggregation and analysis. In the first step, we automatically translate the 500 BLEnD questions into Swedish using ChatGPT's translation API. We then ask a native Swedish speaker to review the data and correct any errors. We noted that, in general, ChatGPT had decent translation quality; however, a few concepts were a bit unclear.

For instance, *What is a popular snack at an amusement park in Sweden?* was translated by ChatGPT to *Vad är ett populärt mellanmål på en nöjespark i Sverige?*, our native speaker corrected that into *Vilket mellanmål är populärt på nöjesfält i Sverige? (T.ex., Gröna Lund eller Liseberg)*. This was one of the questions where the translation sounded anglicized, particularly in the word order, but also in that the cultural knowledge is highly relevant. It is relatively rare to hear the word *nöjespark* or *nöjesfält* in Sweden; since there are few amusement parks, people tend to refer to them by their brand names, such as Gröna Lund, Liseberg or Tivoli. Figure 1 illustrates a few examples from the process and an example of the type of questions that are part of the dataset. Each question belongs to one of 6 categories: Food, Sports, Family, Education, Holidays, Work-life (from the original paper). Each question is given to five different participants to answer.

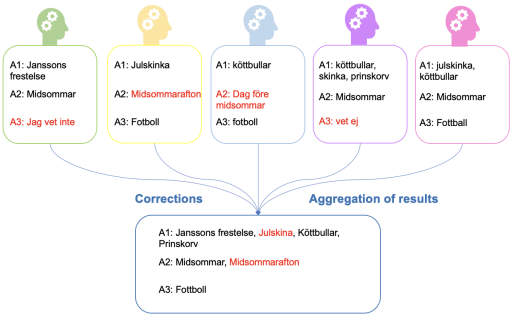

Figure 2: All responses had to be aggregated. For the aggregations, we asked another native speaker who was not among the participants to review all responses. They had to correct any misspellings or correct concepts if needed. They should not provide an answer themselves, but correct and merge the answers from the participants. Translations are all provided as well.

3.1. Response Collection

Once we had the 500 questions correctly translated, we recruited five Swedish participants and provided them with instructions for providing responses. The participants were all paid according to an hourly rate for research assistance in Sweden. The participants did not interact with or influence one another's responses. We collected all responses for finalizing the dataset and analysis.

For each question, participants were required to provide at least one short answer and were allowed to give up to three responses for questions where alternatives were relevant. For questions where participants did not know the answer in the Swedish context, they had the option of giving one of the following answers: not applicable to Sweden", "No specific answer to this question", "I don't know", or other". If participants answered "I don't know", the response was excluded from the aggregation, as our evaluation focuses on comparing model outputs to produced answers. However, we acknowledge that such responses may reflect meaningful uncertainty and could be explored in future work.

At the aggregation stage, one designated reviewer in Sweden examined all collected human responses, removed invalid or nonsensical responses that may have resulted from misunderstandings, and consolidated lexical variants of the same concept (e.g., "go to bed" and "sleep") to ensure accurate vote counts. The reviewer also translated the answers into English. The final Swedish BLEnD dataset, therefore, includes the original Swedish responses, their English translations, consolidated answer groupings, and the final vote counts for each question. Figure 2 shows an

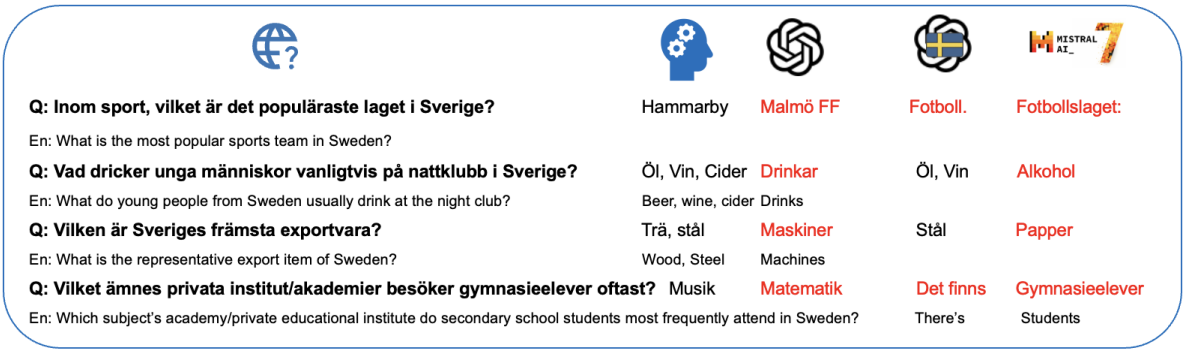

Figure 3: Examples of Human Answers vs. LLMs Generations. In red are all the wrong answers that different models generated.

example of our aggregation scheme. In this case, we have two interesting cases, for question 2, relating to the most common Swedish holiday, all participants agreed that it is "Midsommar", which is the common phrase for "Midsommarafton" (*Midsummer Eve*) in Sweden,[4]. Participants 2 and 3 agree that it is Midsummer Eve; however, they write it differently. Participant 2 uses the Swedish "midsommarafton", whereas Participant 3 uses a phrase to refer to either the same day or the day before (in some regions of Sweden, Midsummer is celebrated over more than one day). Our aggregator keeps only two answers in this case: "Midsommar" and "Midsommarafton".

4. LLM Evaluation

The purpose of the BLEnD benchmark is to evaluate the extent to which large language models (LLMs) encode everyday knowledge. To this end, we generate LLM answers using **GPT-4o-mini**, **GPT-SW3** (Ekgren et al., 2022),[5] and **Mistral 7B** (Jiang et al., 2023) on our dataset. Table 1 reports both exact-match accuracy and cosine similarity accuracy across the datasets. We treat this task as a constrained answer setting rather than open-ended generation, since the models are explicitly instructed to respond with one or two words only. That is, exact-match accuracy provides a simple and interpretable way of measuring whether the model produces the same answers as the participants. We therefore use accuracy as a strict metric, and complement it with cosine similarity to capture semantically similar responses.

Despite the constrained output format (one or two words), the range of plausible answers remains large, meaning that a random baseline would be close to zero. The only model that achieves non-trivial performance is gpt-4o, with correct answers on roughly half of the questions. All other models fail in their responses, which reveals systematic limitations rather than isolated errors. To better understand these shortcomings, we conducted a qualitative analysis focusing on the types of questions that challenge the models.

Figure 3 presents illustrative comparisons between human responses and GPT-4o outputs. In the first example, all participants identified Hammarby, whereas GPT-4o instead produced Malmö FF. This difference is not necessarily random, and the model is not necessarily wrong *per se*. Presumably, this model output reflects a reliance on global frequency patterns rather than contextually grounded everyday knowledge, which is, however, consistent with simple regional variation in what is considered a "popular" team. Malmö may also be more popular overall, given its sporting success, whereas our participants are Stockholm-based and their responses likely reflect that perspective. The second and third examples further expose this limitation. GPT-4o generated Drinkar and Maskiner ("drinks" and "machines"), responses that are lexically plausible but pragmatically misaligned. These answers are not semantically incoherent; rather, they seem to reflect a failure to capture culturally situated meaning, but could also be due to alternative interpretations of the question or more general pragmatic ambiguity in how the question is understood. Across multiple instances, the models produce outputs that are superficially compatible with the question but lack the implicit social or contextual grounding that human respondents readily apply. Taken together, these findings suggest that LLMs rely heavily on surface-level co-occurrence

[4]Translations and explanations: "Janssons frestelse" is a potato-based dish; "Julskina" (*Christmas Ham*); "köttbullar" (*meatballs*); "prinskorv" is a type of sausage; "Midsommar" is a Swedish traditional holiday; "Midsommarafton" (*Midsummer Eve*); "Dag före Midsommar" (*Day before Midsummer*)

[5]https://huggingface.co/AI-Sweden-Models/gpt-sw3-1.3b-instruct

Model	Corr. (strict)	Corr./Total	Cosine Similarity	Corr_cosine/Total
gpt-4o	**50.80%**	254/500	**51.60%**	258/500
gpt-sw3-1.3b-instruct	15.80%	79/500	17.80%	89/500
Mistral7B	0.80%	4/500	11.40%	57/500

Table 1: Strict lexical accuracy and cosine similarity against BLEnD gold answers for our evaluation dataset. We used the same prompt with all models: **Svara med ett eller två ord på svenska. Endast ord, ingen förklaring, ingen punkt** which translates to: Answer with one or two words in Swedish, no explanation, no full stop.

.

statistics and global prominence signals. While this strategy is often sufficient for general factual or associative knowledge, it breaks down when questions require culturally embedded, community-specific, or pragmatically constrained understanding. The BLEnD benchmark thus exposes a gap between distributional competence and culturally grounded everyday knowledge.

We leave comparisons to other BLEnD languages for future work, seeing as it would require a fairly well-controlled setup across languages, which would have to include (almost) identical prompts, decoding settings, model versions, and answer aggregation procedures. For example, even small differences in prompting (e.g., constraining answers to one or two words), or in how human responses are aggregated, could affect the outcome.

5. Conclusion

In this paper, we extend the everyday knowledge benchmark, BLEnD, to include Swedish. We also evaluate a few LLMs on Swedish data and show that aggregate accuracy masks systematic weaknesses: while models perform well on roughly half of the questions, qualitative analysis reveals recurring failures on items requiring culturally situated reasoning. These errors are typically not lexically implausible but pragmatically and culturally misaligned, suggesting a reliance on distributional prominence rather than grounded understanding. BLEnD thus highlights a gap between surface-level linguistic competence and culturally embedded everyday knowledge. We hope this benchmark encourages more fine-grained evaluation practices that account for cultural grounding.

6. Limitations

In this paper, we did not introduce a new dataset of culturally grounded Swedish knowledge; rather, we extended the BLEnD benchmark to Swedish. The goal was to examine whether LLMs maintain culturally informed reasoning when applied to a low- to medium-resourced language, or whether this capability degrades outside high-resource settings.

We view this work as a first step. Benchmarks shape model development: as systems are trained on increasingly diverse data and evaluated on more targeted benchmarks, they adapt to these evaluation signals. Although benchmarks risk becoming outdated as models improve, this does not argue against creating them. On the contrary, continuous development of culturally and pragmatically challenging benchmarks is essential for stress-testing emerging technologies and tracking their limitations over time.

Finally, we acknowledge the limitations of our study's scope. The benchmark captures only a narrow slice of Swedish cultural knowledge and cannot represent the diversity of perspectives across Sweden. In particular, although three of the five participants did not grow up in the Stockholm area, two did, and all five currently live there. Future extensions should aim to achieve broader geographic and demographic coverage to better reflect cultural variation across Sweden.

Acknowledgements

We wish to thank the five participants who provided answers. We also thank the three anonymous reviewers, whose comments were indeed constructive and helpful.

7. Ethical Considerations

In this work, we paid participants from different backgrounds to answer questions about Swedish culture. No personal or sensitive data were collected. We do not claim to capture the nuances of an entire culture in a single dataset. The primary purpose of the dataset is to evaluate whether LLMs can accurately predict certain aspects of cultural knowledge. Our goal is to provide a starting point focused on a dataset of everyday concepts. Another aspect linked to large language models is that this dataset will most likely be part of the next training data for new models, which means that we should be careful about generalizations in the future. Lastly, we used ChatGPT to fix grammar and spelling.

References

Junia Anacleto, Henry Lieberman, Marie Tsutsumi, Vânia Neris, Aparecido Carvalho, Jose Espinosa, Muriel Godoi, and Silvia Zem-Mascarenhas. 2006. Can common sense uncover cultural differences in computer applications? In *Artificial Intelligence in Theory and Practice. IFIP AI 2006. IFIP International Federation for Information Processing*, pages 1–10, Boston, MA. Springer US.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Niclas Hertzberg and Anna Lokrantz. 2024. MedQA-SWE - a clinical question & answer dataset for Swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186, Torino, Italia. ELRA and ICCL.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jenny Kunz. 2025. A diagnostic benchmark for sweden-related factual knowledge.

Stella Lundqvist. 2025. Do large language models and humans follow similar learning stages?: Assessing GPT-2's order of Swedish grammar acquisition within the Processability Theory framework. Master's thesis, Uppsala University.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tobias Norlund, Tim Isbister, Amaru Cuba Gyllensten, Paul Dos Santos, Danila Petrelli, Ariel Ekgren, and Magnus Sahlgren. 2024. Sweb: A large web dataset for the scandinavian languages.

Nedjma Ousidhoum, Shamsuddeen Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Ahmad, Sanchit Ahuja, Alham Aji, Vladimir Araujo, Abinew Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine Kock, Genet Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Yimam, and Saif Mohammad. 2024. SemRel2024: A collection of semantic textual relatedness datasets for 13 languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2512–2530, Bangkok, Thailand. Association for Computational Linguistics.

Johan Sjons, Fredrik Heinat, and Murathan Kurfali. 2026. The swedish benchmark of linguistic minimal pairs. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain. To appear.