Understanding Multilingual ASR Systems: The Role of Language Families and Typological Features in Seamless and Whisper

Simon Gonzalez, Tao Hoang, Maria Myung-Hee Kim,
Bradley Donnelly, Jennifer Biggs, and Tim Cawley
Defence Science and Technology Group
simon.gonzalez@defence.gov.au, tao.hoang@defence.gov.au, myung.kim@defence.gov.au,
bradley.donnelly2@defence.gov.au, jennifer.biggs@defence.gov.au, tim.cawley@defence.gov.au

Abstract

This study investigates the extent to which linguistic typology influences the performance of two automatic speech recognition (ASR) systems across diverse language families. Using the FLEURS corpus and typological features from the World Atlas of Language Structures (WALS), we analysed 40 languages grouped by phonological, morphological, syntactic, and semantic domains. We evaluated two state-of-the-art multilingual ASR systems, Whisper and Seamless, to examine how their performance, measured by word error rate (WER), correlates with linguistic structures. Random Forests and Mixed Effects Models were used to quantify feature impact and statistical significance. Results reveal that while both systems leverage typological patterns, they differ in their sensitivity to specific domains. Our findings highlight how structural and functional linguistic features shape ASR performance, offering insights into model generalisability and typology-aware system development.

1 Introduction

Recent advances in multilingual automatic speech recognition (ASR) have attracted growing attention to how models process and generalise across languages (Yadav and Sitaram, 2022; Heigold et al., 2013; Li et al., 2025). Much of the current research on multilingual ASR focuses on model architecture and optimisation techniques, especially for enhancing cross-lingual transfer capabilities (Anidjar et al., 2023; Liu et al., 2021). For example, Huang et al. (2024) propose language embedding methods to improve ASR performance on unseen languages, highlighting enhancements in model design and parameter sharing. While these approaches have achieved measurable performance gains, they tend to prioritise engineering solutions over linguistically grounded interpretations of ASR behaviour.

In parallel with these system-driven approaches, a complementary line of research examines the relationship between linguistic properties and ASR performance. Prior work has demonstrated that leveraging linguistic similarity enables multilingual ASR models to generalise to languages not included in their training data. Phonetic typology, in particular, has proven to be an effective predictor of multilingual ASR performance on unseen languages. This effectiveness is driven by the model's ability to extract phonetic patterns from training languages and apply them to typologically related ones, thereby improving recognition accuracy (Zellou and Lahrouchi, 2024). However, Feng et al. (2021) identified key limitations in modelling phonotactics across different languages in multilingual ASR systems, suggesting that generalising phonotactic patterns across languages may not always lead to performance gains.

Semantic similarity between languages has also been explored as a resource for improving multilingual ASR. Anidjar et al. (2023), for instance, developed a semantic dataset and applied a pretrained speech representation model [Wav2Vec 2.0 (Baevski et al., 2020)] to examine how shared semantic features can facilitate cross-lingual recognition. However, their study did not examine variation across language families or engage deeply with linguistic typology. These findings suggest that evaluating multilingual ASR through a linguistic feature framework can provide deeper insight than analysing individual languages in isolation. By examining which aspects of language structure a model attends to, i.e., whether phonological, morphological, syntactic, or semantic, we can gain a clearer understanding of multilingual ASR performance and generalisation.

This perspective is supported by Ferrand et al. (2024), who examined the robustness of neural ASR systems on polysynthetic languages and highlighted persistent challenges when handling morphologically complex languages. Such findings underscore the importance of evaluating multilingual

ASR models not only in terms of language-specific performance, but also in their ability to capture cross-linguistic generalisations.

This study provides a typologically-informed analysis of multilingual ASR performance. Using the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2005), we represent 40 languages through 168 typological features grouped into four linguistic domains: phonology, morphology, syntax, and semantics. These domains reflect a structural-functional continuum in language, from sound patterns to meaning-based features. We then analyse how performance, measured by Word Error Rate (WER), correlates with these domains across two state-of-the-art multilingual ASR systems: Whisper and Seamless.

We employ Random Forest models (Breiman, 2001) to estimate the relative importance of each linguistic feature and cluster them by impact level. In addition, we incorporate language family membership into the analysis to examine whether ASR performance patterns align with genealogical relationships. Mixed Effect Models (Silveira et al., 2023) are used to assess the significance of domain-level effects. By comparing Whisper and Seamless, we assess whether architectural differences lead to distinct patterns of sensitivity to typological features and language family structure.

Our results contribute to understanding whether ASR systems rely on language-specific learning or can generalise from structural patterns shared across typologically related languages. Our findings offer a deeper understanding of how structural and functional linguistic features shape ASR outcomes. This contributes to the development of more interpretable and equitable multilingual ASR systems.

2 Related Work

Previous research on ASR performance has examined a range of languages, but the scope and focus of these studies vary considerably. Some investigations have focused on relatively smaller sets of languages, often selected for practical reasons such as data availability or coverage in existing benchmarks (Attanasio et al., 2024; Gonzalez et al., 2024; Heigold et al., 2013). While such studies provide valuable insights into system performance in specific contexts, their findings are limited in their ability to capture typologically broader trends.

Other work has addressed larger, more diverse language sets, including those used in multilingual benchmarks and shared tasks. These studies often report descriptive performance metrics across languages, but without explicitly incorporating linguistic family membership or typological features into the analysis (Pratap et al., 2020). As a result, they are less able to identify which aspects of linguistic structure most directly influence recognition accuracy, or whether systems exhibit systematic behaviour across related languages.

The present study extends this literature by adopting a typology-informed approach that integrates linguistic and structural information into ASR evaluation. By grounding the analysis in linguistic theory and drawing on established typological resources, we aim to move beyond descriptive comparisons and towards a more principled understanding of how linguistic diversity shapes ASR performance.

3 Methodology

To investigate the role of linguistic structure in multilingual ASR, we examine how four core linguistic domains, i.e., phonology, morphology, syntax, and semantics, interact with ASR models in shaping WER outcomes. We define each domain by a set of typological features sourced from the World Atlas of Language Structures (WALS), grouped by language family. We analyse how these features correlate with WER across languages and use Random Forest (RF) regression to quantify their relative importance. Mixed Effects Models (MEMs) are then applied to test the statistical significance of each domain. We compare the performance of two multilingual ASR systems, namely Whisper and Seamless, by examining how each model's WER correlates with typological features across linguistic domains and language families.

3.1 FLEURS Dataset

The speech data for this study comes from the FLEURS dataset (Conneau et al., 2022), which contains transcribed and translated speech recordings of read sentences across a wide range of languages. Forty languages were selected to maximise diversity in language family, phonological systems, and grammatical structures. FLEURS is particularly well suited to this investigation because it offers balanced and comparable data across languages, allowing for controlled cross-linguistic comparisons of ASR performance. The dataset has also been

used extensively in prior ASR and multilingual evaluation research, making it a reliable benchmark for typologically informed ASR evaluation across systems and languages.

3.2 Typological Features Dataset

To capture the linguistic characteristics of each language, we used the WALS dataset, which provides a comprehensive set of typological features compiled over decades of empirical linguistic research, covering a wide range of linguistic aspects. Each language in our study was linked with its language family classification and a set of typological features, which allowed us to quantify structural similarities and differences.

For analytical clarity, the 168 WALS features selected for this study were grouped into four major linguistic domains: phonology, morphology, syntax, and semantics. While these domains are presented separately, they represent interrelated dimensions of linguistic structure that lie along a continuum, with no absolute boundaries between them. At one end of this spectrum, phonology and morphology are closely linked through their grounding in the sound structure of language. At the other, syntax and semantics reflect functional and meaning-based aspects of linguistic organisation. This continuum provides the conceptual framework for interpreting ASR results, with phonology situated at the structural end and semantics at the functional end of the spectrum.

The domains are defined as follows:

Phonology (20 features): The organisation of the sound system, including phoneme inventories and phonotactic constraints.

Morphology (60 features): The internal structure of words, including inflectional patterns, and structural complexity.

Syntax (80 features): The arrangement of words and constituents, including word order patterns and clause structures.

Semantics (8 features): Features related to meaning, including lexical categories, and meaning distinctions encoded in the language.

3.3 ASR Systems

We evaluated two widely used multilingual ASR systems: Whisper, developed by OpenAI (Radford et al., 2022) and Seamless, developed by Meta AI (Barrault et al., 2023). They represent state-of-the-art approaches to speech recognition across a broad range of languages. For our analysis, we used Whisper Large v2 and SeamlessM4T v2 model variants, which demonstrate strong performance on multilingual speech-to-text tasks but differ substantially in design. Evaluating these systems side by side allows us to investigate whether differences in model architecture and training manifest in distinct patterns of typology-related performance variation. Table 1 summarises key distinctions between the two systems.

Aspect	Seamless	Whisper
Tasks	Multilingual Speech TTS/ Text	Multilingual STT
Languages	100	97
Architecture	Transformer: Speech/ Text Encoder Text Decoder Text-to-Unit Vocoder	Transformer: Speech Encode + Text Decoder
Data	Supervised 496K hours: Speech-Text Pairs, Text-Text Pairs. Self-Supervised 4.5 million hours: Speech only	Supervised 680K hours: Speech-Text Pairs.
Focus	Cross-lingual Speech-Text Alignment. More focus on lower-resource languages	Robust multilingual ASR. Biased towards higher-resource languages

Table 1: System comparison between Seamless and Whisper.

3.4 Typology-Based Evaluation Methods

Random Forest: To investigate whether ASR performance varies with respect to language family membership and associated typological features, we employed a Random Forest analysis. This machine learning approach was used to evaluate typological features collectively and estimate their relative importance in explaining WER variation across languages, enabling us to assess how language family membership and typological characteristics relate to ASR performance.

Based on the variable importance scores derived from the Random Forest, we applied cluster analysis to group features into three levels of impact

on ASR accuracy, namely, low, medium, and high. This grouping facilitates interpretation by highlighting which linguistic factors most strongly explain the variability in WER. The proportion of features from each linguistic domain within these clusters serves as an indicator of their relative influence. For example, a higher concentration of phonological and morphological features in the top importance cluster suggests that the performance of the ASR system is closely tied to properties of the language sound system. On the other hand, if semantic and syntactic features predominate in the top cluster, this implies that the system errors are more related to language meaning and structure.

Conceptually, these four linguistic domains can be arranged along a continuum from the sound system (phonology) to meaning (semantics), with morphology bridging phonology and syntax, and syntax bridging morphology and semantics. Therefore, errors associated with phonology and morphology tend to reflect challenges in processing acoustic and sub-word sound patterns, whereas those associated with syntax and semantics point toward difficulties in handling structural and meaning-related aspects of language. This approach provides a nuanced understanding of the typological factors influencing multilingual ASR systems and reveals distinct patterns in how different models leverage linguistic information across languages.

Mixed Effects Models: While the Random Forest analysis identifies the most important linguistic features associated with ASR performance, it does not provide information about the statistical significance of their effects. To address this, we employed Mixed Effects Models to examine how each feature impacts WER and to assess the role of language family membership more explicitly. For each linguistic feature, we fitted two linear MEMs with language treated as a random effect to account for variability across individual languages. The first model included the linguistic feature and language family as the fixed explanatory variables, while the second model served as a baseline without the language family. With this, we do not only have insight into the importance of the feature for WER but also whether this behaviour changed based on the language family.

We then conducted likelihood ratio tests (ANOVA (Girden, 1992)) to compare the two models and determine whether the inclusion of language family significantly improved the fit of the model. This approach allowed us to evaluate the statistical significance of each feature contribution to explaining WER variation, with a particular focus on the effect of language family membership. Only features showing significant differences in the ANOVA tests are discussed further, providing a focused interpretation of the most impactful typological factors influencing ASR performance.

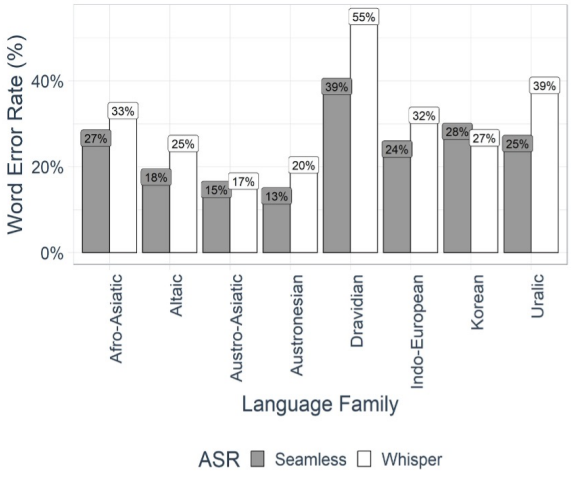

Figure 1: WER by Language Family and ASR System

4 Results

Initial WERs show that Seamless outperforms Whisper on average, achieving 25% WER compared to 33% for Whisper across all languages. At the individual language level (see Appendix A), only Korean, Serbian and Swedish show higher WER for Seamless, whereas in all the other languages, Seamless produces lower WERs. When grouped by language family, patterns of performance differ. Figure 1 presents average WER across language families for each ASR system.

WER ranges from 13% for Austronesian language family with Seamless and 55% for Dravidian language with Whisper. To examine patterns of internal grouping among language families, we conducted a cluster analysis using WER estimates derived from the MEMs. Dendrograms were generated to visualise the hierarchical relationships between language families on ASR performance (see Figure 2).

Dendrograms show that both Whisper and Seamless produced similar clustering patterns, grouping Uralic, Indo-European, and Afroasiatic together on one side and separating the Dravidian family on the other, while also clustering Austronesian and Austroasiatic internally, which is a result con-

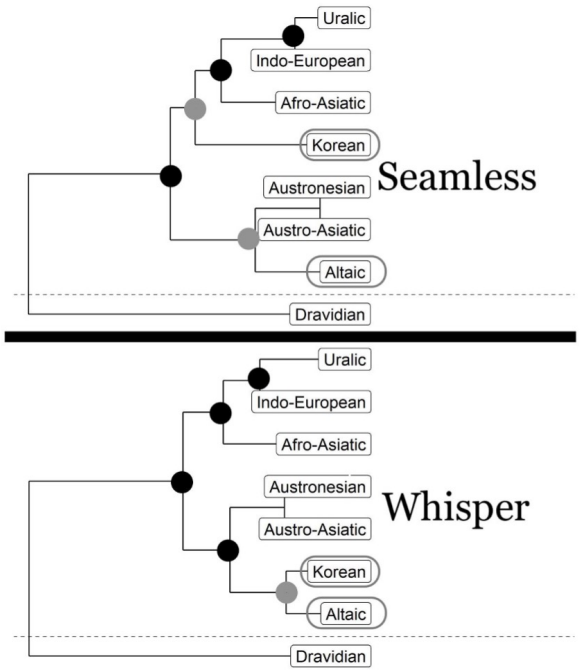

Figure 2: Dendrograms of Language Family by WER Estimates

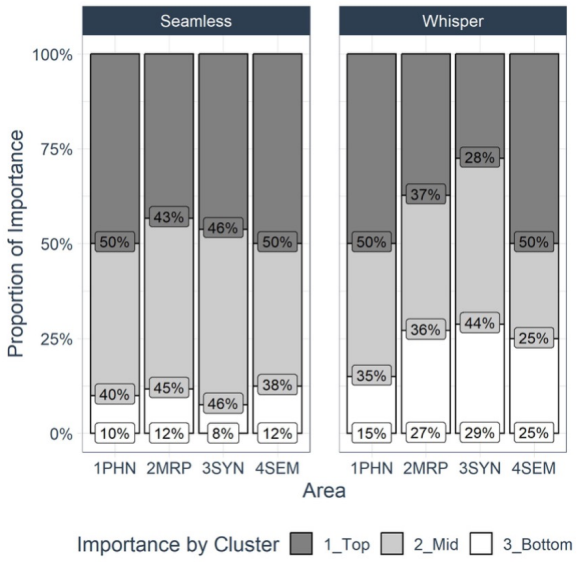

Figure 3: Random Forest Feature Importance Categories for each ASR System

sistent with known phonological and grammatical parallels (Dunn et al., 2008; Hammarström et al., 2022).

The key difference lays in the treatment of Korean and Altaic. Whisper grouped them near Austronesian/Austroasiatic, whereas Seamless separated them, positioning Korean closer to Uralic/Indo-European/Afroasiatic families. Whisper's clustering aligns more closely with linguistic research highlighting phonological and morphological similarities between Korean and Altaic (Robbeets, 2005; Janhunen, 2007; Robbeets and Savelyev, 2020).

4.1 Random Forest Results

The Random Forest analysis revealed distinct patterns in the relationship between linguistic features and ASR performance for the two systems under study. Figure 3 shows the importance values for both ASR systems. All features are grouped into the four linguistic Areas and the cluster of importance: Top, Mid, and Bottom importance.

For Seamless, the importance of features has a very similar distribution across the different areas, with the Syntactical features (92%) encompassing most proportions of importance when both Top (46%) and Mid (46%) levels of importance are considered. This suggests that for Seamless, all the features play a similar role of importance when transcribing speech from the audio signal. This pattern is in line with human languages where the understanding of communication is maximised using multiple cues simultaneously and not only focusing on one linguistic feature (Ding et al., 2023).

When compared to Seamless, Whisper's highest areas of importance are also Phonology and Semantics, both with 50% of features in the Top tier importance. However, the difference lies in Whisper's treatment of Morphological (37%) and Syntactical (28%) features, which are both lower than Seamless. This suggests that on the one hand, both systems rely strongly in the sound structure of language to achieve accuracy at the level of language meaning. On the other hand, they have different ways in how they deal with the intermediary linguistic features within the structure-meaning continuum. Whisper prioritises morphological features rather than syntactical features, suggesting that errors in Whisper are more closely tied to the structural aspects of language, particularly in word forms, than to word order and clause structures.

Combining these results, both systems showed the role of phonology and semantics in predicting recognition errors, but Seamless exhibited a greater emphasis on a holistic approach, while Whisper was more sensitive to form-related features, particularly those tied to morphological complexity.

4.2 Mixed Effects Models Results

The MEMs provided further nuance to these observations. For this analysis, we only present those

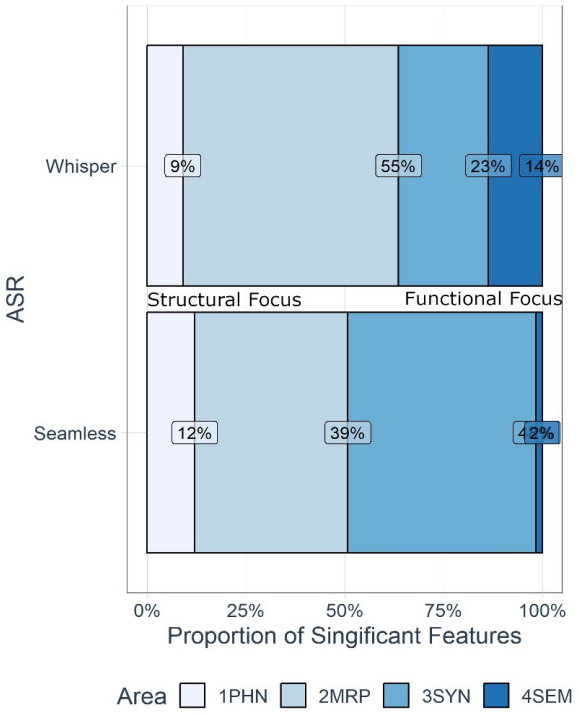

Figure 4: Proportion Distribution for all Significant Linguistic Features for each ASR System

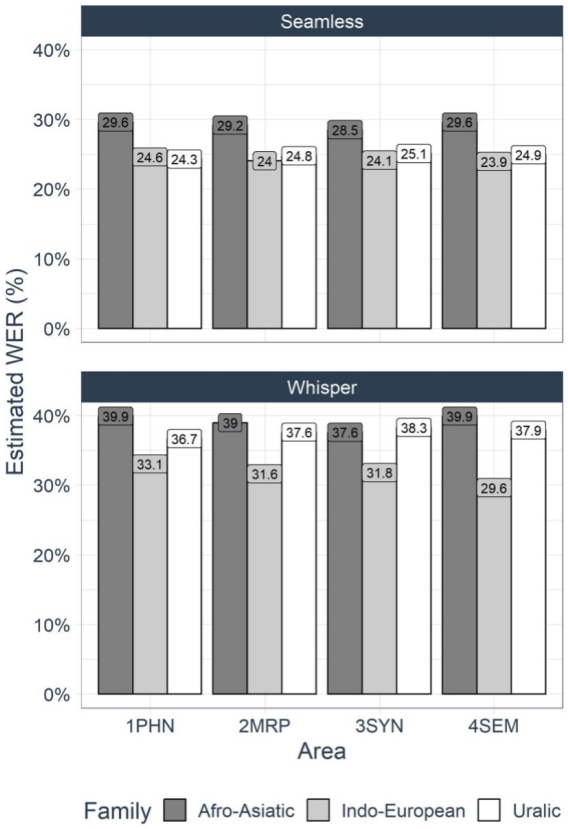

Figure 5: Interaction between Language Family and Linguistic Area

cases where there is a significant difference for language family affiliation for a given feature. All significant comparisons are grouped into the four linguistic Areas. For each ASR, all features are grouped to total 100%. The individual percentages for each represent the focus of features that produce the most significant differences. Figure 4 shows the results for both ASR systems across the four areas.

For Whisper, 64% of the significant predictors of WER fell towards the more Structural Focus of features (the combined Phonological 9% and Morphological 55% domains), with Morphological features accounting for the majority of this influence. Phonological features played a secondary role, but their presence alongside Morphology points to Whisper relying on the formal/structural properties of language in shaping recognition accuracy.

In contrast, Seamless shows a more balanced weight between Functional and Structural Focus. However, its stronger area, Syntax (49%), adds more weight towards the Functional Focus (50% when combining Syntax (48%) and Semantics (2%)). This aligns with the Random Forest results in indicating that Seamless is strongly affected by the functional organisation of language, mainly on the Syntactic structure, than by purely structural characteristics offered by Phonology (12%) and Morphology (39%).

An interaction analysis of language family and linguistic area was then performed to assess how this observed pattern was consistent across language families. The analysis revealed that only three language families displayed greater differences between areas, while the remaining families showed no appreciable variation of accuracy between domains. Figure 5 shows these patterns.

For the Seamless model, the Afro-Asiatic family consistently exhibits a higher error rate than the Indo-European and Uralic families, which cluster together and display comparatively smaller area differences, with a maximum difference of 1.1% between domains. In contrast, the Whisper model shows that the errors in the Uralic family pattern more closely to those of the Afro-Asiatic group, and the domain differences are larger than in Seamless, with a maximum difference of 3.5%. This is an indication that ASR accuracy is not only model-dependent but can also be affected by the interaction between language family and linguistic domain, with Whisper exhibiting greater discrep-

ancies in accuracy across the four areas.

4.3 Combining Results from RFs and MEMs

When comparing the RF and MEM results, clear differences emerge in the role of Phonology and Semantics. The RF analysis positions the domains as the most influential predictors of accuracy, whereas the MEM results attribute comparatively lower proportions of explained variance to them in both systems. This divergence suggests that ASR performance is highly sensitive to a small subset of phonological and semantic features, which, although few, have disproportionate predictive power.

From a theoretical perspective, this pattern may indicate that ASR systems are optimised to exploit high-impact phonological and semantic cues only when they are particularly salient, rather than relying on them broadly. By contrast, morphological and syntactic domains appear to contribute through a larger pool of features, reflecting a more distributed and complex influence on WER variability.

This could reflect the computational strategies of ASR architectures, which may prioritise morphosyntactic structures for robust performance across diverse linguistic inputs, while reserving phonological and semantic information for resolving specific, high-information contexts. Such findings underscore the need to consider not only the impact of feature effects but also their distribution across linguistic domains when evaluating the linguistic adequacy of ASR systems.

5 Discussion

The analyses presented in this study offer valuable insights into the linguistic factors shaping ASR performance across diverse language families. Our findings demonstrate that ASR accuracy is indeed influenced by the language family to which a language belongs, although the strength of this effect varies. Some language families exhibit consistently higher or lower WERs, indicating that not all linguistic families present the same challenges to ASR systems. These results suggest that multilingual ASR systems are sensitive not only surface-level phonetic variation but also deeper structural features. The dendrograms broadly align with genealogical groupings, implying that the models may implicitly learn linguistic structure. A key divergence is found in how Korean and Altaic are clustered, which reflects an area of ongoing debate in historical linguistics, showing how contested language relationships can surface even in computational models.

However, ASR performance cannot be attributed solely to language family. Rather, it correlates with a complex interplay of typological features spanning phonology, morphology, syntax, and semantics. Our results suggest that multilingual ASR models leverage shared linguistic patterns that transcend individual languages, supporting a form of generalisation similar to cross-linguistic transfer. This mirrors how humans process language and has parallels with Large Language Models, which generalise learned knowledge across related languages by internalising abstract linguistic structures. Moreover, our findings reveal that Seamless and Whisper adopt different strategies of focus. Seamless takes a more holistic approach, leveraging all domains with a strong emphasis on language function. In contrast, Whisper places greater weight on structural aspects, particularly phonology and morphology. Notably, Seamless consistently outperforms Whisper across the diverse language set, suggesting that prioritising functional aspects of language, those related to meaning and grammatical function, may lead to improved ASR performance.

The fact that Seamless outperforms Whisper in WER is not merely a matter of engineering efficiency but reflects a deeper linguistic orientation in its training design. As described in Barrault et al. (2023), the model was trained by aligning semantically similar languages, effectively grounding its internal representations in shared meaning structures rather than treating each language as an isolated system. This strategy is aligned with long-standing linguistic theories that emphasise the role of universals and cross-linguistic transfer in shaping communicative systems (Chomsky, 1965; Greenberg, 1963; Odlin, 1989; Ruder et al., 2019). By learning from clusters of related languages, Seamless is able to capture semantic and syntactic patterns that generalise beyond surface variation, enabling a more holistic and linguistically informed architecture. In contrast, Whisper's narrower focus on uniform data coverage leads to strong robustness but lacks the same degree of linguistic depth. The lower WER achieved by Seamless can thus be interpreted as a result of its training paradigm, which mirrors how humans exploit linguistic similarity and shared structures across languages to facilitate understanding.

Our findings indicate that morphological and syntactic features often show stronger influence on ASR accuracy than phonological structure, particularly for end-to-end models trained on limited data. This aligns with established linguistic intuition that speech sound inventories, phonotactic constraints, and morphological richness directly affect acoustic and lexical modelling stages of ASR. The cross-system comparison highlighted that while different ASR architectures respond similarly to broad linguistic challenges, some models are more resilient to specific domains of complexity. These insights suggest the potential for linguistically-informed model selection or fine-tuning strategies.

6 Conclusions

This study examined how linguistic features drawn from WALS and spanning multiple language families can explain ASR performance across a diverse set of languages in the FLEURS dataset. By grouping these features into four core domains (phonology, morphology, syntax, and semantics), and comparing results across multiple ASR systems, we identified both language-specific and typology-driven effects on WER.

This study underscores the importance of integrating linguistic typology into ASR research and development. By anticipating which languages and features are likely to challenge a given ASR system, developers can tailor training, data augmentation, and evaluation methods more effectively. Future research should expand beyond WALS to incorporate prosodic, pragmatic, and discourse-level features, and explore hybrid architectures that explicitly account for linguistic diversity.

By incorporating a typologically diverse set of 40 languages across language families, this study advances a more comprehensive understanding of language behaviour and ASR performance across the world's linguistic diversity. The present investigation can also provide users with a detailed, language-family-specific overview of each model's performance. By delineating the potential failure points for each system in relation to linguistic characteristics, users can more accurately select or refine an ASR solution that aligns with the linguistic characteristics of their target language.

7 Limitations and Future Work

This study presents several limitations that should be addressed in future research. First, the analysis was restricted to two ASR models, which may limit the generalisability of the findings. Expanding the scope to include a broader range of ASR systems would provide a more comprehensive understanding of performance variation across models. Second, the investigation focused exclusively on pre-trained ASR systems. Incorporating fine-tuned models in subsequent studies would enable an examination of how domain-specific adaptation influences performance across different linguistic contexts.

Also, although the dataset used is relatively extensive, its coverage of language families remains incomplete. Extending the dataset to include additional languages, particularly from underrepresented linguistic families, would enhance the robustness of the observed patterns and strengthen the conclusions drawn from the analysis. Finally, the speech style examined here was read speech. To ensure that the findings generalise to real-world scenarios, future investigations should evaluate the ASR models on more natural, spontaneous speech, such as conversational and unprompted utterances. This will help identify performance gaps that arise under less controlled, more varied linguistic conditions.

Acknowledgments

We gratefully acknowledge the comments from Hayden Ooi, as well as the three anonymous reviewers. Their insightful observations and suggestions on the earlier draft enhanced the quality of this paper, and we are deeply appreciative of their invaluable input.

References

O. H. Anidjar, R. Yozevitch, N. Bigon, N. Abdalla, B. Myara, and R. Marbel. 2023. Crossing language identification: Multilingual asr framework based on semantic dataset creation & wav2vec 2.0. *Machine Learning with Applications*, 13:100489.

G. Attanasio, B. Savoldi, D. Fucci, and D. Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 21318–21340. Association for Computational Linguistics.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

L. Barrault, Y. A. Chung, M. C. Meglioli, D. Dale, N. Dong, P. A. Duquenne, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

A. Conneau, Y. Bian, P. Rivière, and 1 others. 2022. Flores-200: An evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2022)*.

N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, and M. Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Matthew S. Dryer and Martin Haspelmath, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford, UK.

S. Feng, P. Żelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak. 2021. How phonotactics affect multilingual and zero-shot asr performance. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7238–7242. IEEE.

É. Le Ferrand, Z. Liu, A. Arppe, and E. Prud'hommeaux. 2024. Are modern neural asr architectures robust for polysynthetic languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2953–2963. Association for Computational Linguistics.

E. R. Girden. 1992. *ANOVA: Repeated Measures*. Sage Publications, Inc.

S. Gonzalez and 1 others. 2024. Extending asr systems error measurements: Reporting lexical and grammatical errors. In *Proceedings of the Nineteenth Australasian International Conference on Speech Science and Technology*.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.

G. Heigold, A. W. Senior, M. A. Ranzato, and K. Yang. 2013. An empirical study of learning rates in deep neural networks for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pages 6724–6728. IEEE.

Shao-Syuan Huang, Kuan-Po Huang, Andy T. Liu, and Hung yi Lee. 2024. Enhancing multilingual asr for unseen languages via language embedding modeling. *Preprint*, arXiv:2412.16474.

J. Janhunen. 2007. Typological expansion in the ural-altaic belt. *Incontri Linguistici*, 30:71–83.

J. Li, Y. Shao, J. Zhuo, C. Li, L. Tang, D. Yu, and Y. Qian. 2025. Efficient multilingual asr finetuning via lora language experts. *arXiv preprint arXiv:2506.21555*.

D. Liu, J. Xu, P. Zhang, and Y. Yan. 2021. A unified system for multilingual speech recognition and language identification. *Speech Communication*, 127:17–28.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK.

V. Pratap, A. Sriram, P. Tomasello, A. Y. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert. 2020. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. In *Interspeech / arXiv*.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. Robust speech recognition via large-scale weak supervision (whisper). Technical report, OpenAI.

M. Robbeets and A. Savelyev, editors. 2020. *The Oxford Guide to the Transeurasian Languages*. Oxford University Press, Oxford, UK.

M. I. Robbeets. 2005. *Is Japanese related to Korean, Tungusic, Mongolic, and Turkic?*, volume 64 of *Turcologica*. Harrassowitz, Wiesbaden.

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18. Association for Computational Linguistics.

L. T. Y. D. Silveira, J. C. Ferreira, and C. M. Patino. 2023. Mixed-effects model: A useful statistical tool for longitudinal and cluster studies. *Jornal Brasileiro de Pneumologia*.

H. Yadav and S. Sitaram. 2022. A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2202.12576*.

G. Zellou and M. Lahrouchi. 2024. Linguistic disparities in cross-language automatic speech recognition transfer from arabic to tashlhiyt. *Scientific Reports*, 14(1):313.

Appendices

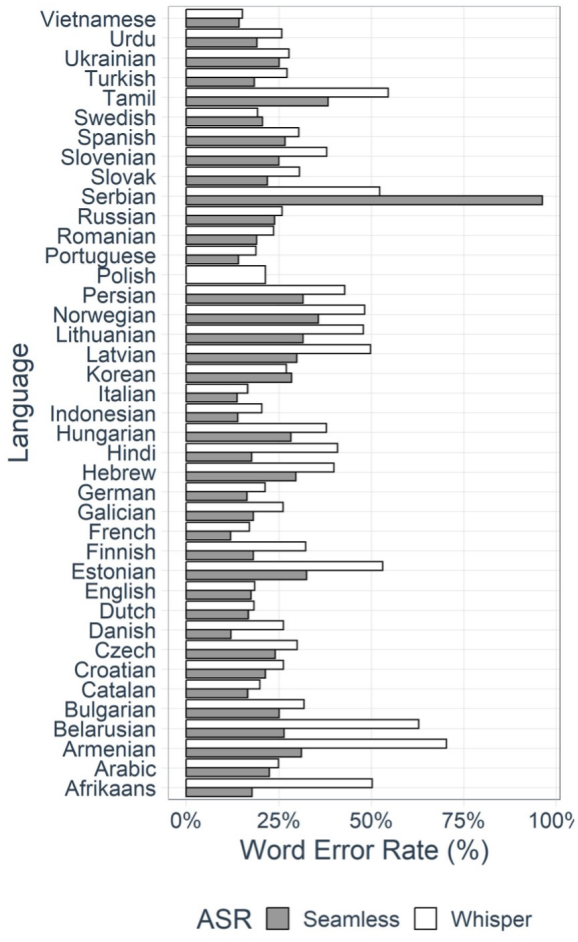

Figure 6: WER For all languages by ASR System

Language	**Files**	**Words**
Afrikaans	1032	23067
Arabic	2104	38164
Armenian	3053	56660
Belarusian	2433	48809
Bulgarian	2973	61026
Catalan	2300	52023
Croatian	3461	65603
Czech	2811	51741
Danish	2465	49695
Dutch	2918	63293
English	2602	54782
Estonian	2501	39596
Finnish	2704	40998
French	3193	76392
Galician	2175	48832
German	2987	61696
Hebrew	3242	54516
Hindi	2120	51062
Hungarian	3095	55937
Indonesian	2579	48837
Italian	3030	68815
Korean	2307	33112
Latvian	2110	36165
Lithuanian	2937	49134
Norwegian	3167	65171
Persian	3101	69276
Portuguese	2793	61693
Romanian	2891	65017
Russian	2562	48365
Serbian	2944	57987
Slovak	1957	35652
Slovenian	2512	48239
Spanish	2796	69137
Swedish	2385	46013
Tamil	2367	37703
Turkish	2526	41336
Ukrainian	2810	51000
Urdu	2109	55980
Vietnamese	2994	86327

Table 2: Number of files and words per language in the dataset.

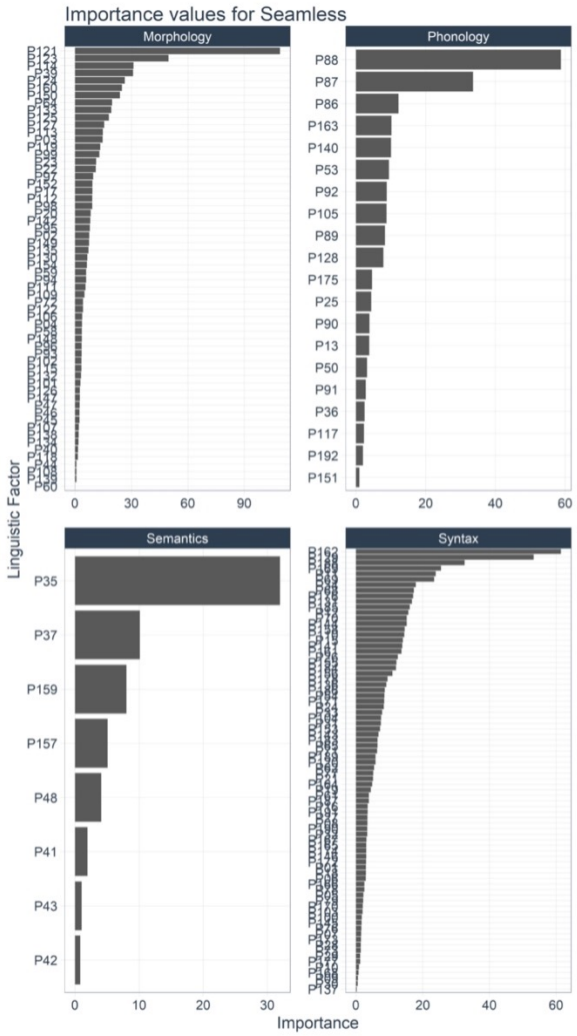

Figure 7: Importance values for Seamless broken down by linguistic domain

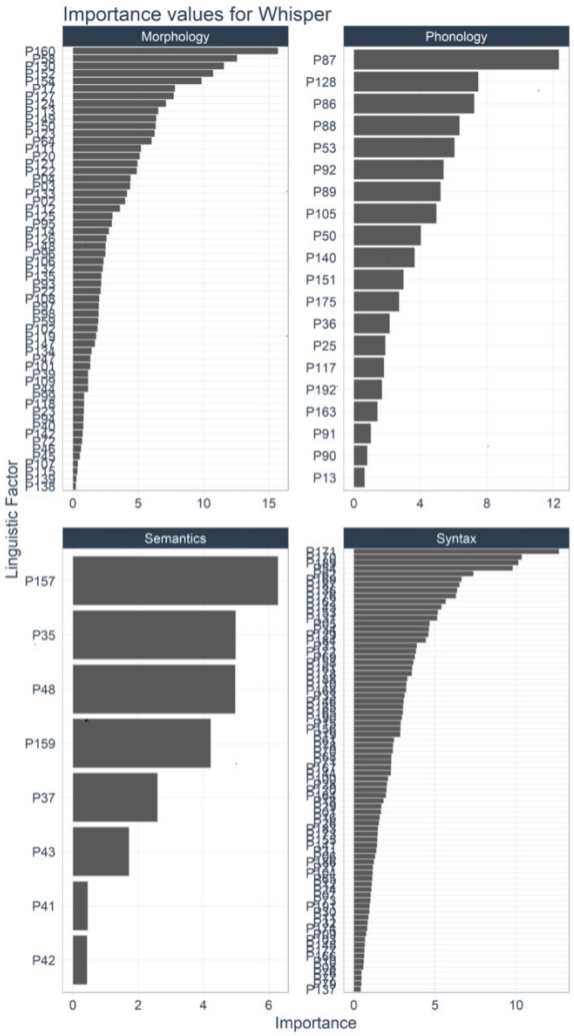

Figure 8: Importance values for Whisper broken down by linguistic domain