Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(12), 554-570
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

# Quantifying Language Disparities: A Distance-Based Predictive Model Using Linguistic Tree Structures

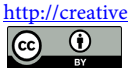
**Akshata Chavan, Nicole Lee, Rezza Moieni, Mary Legrand**

Cultural Infusion Pty Ltd., Melbourne, Australia
Email: akshata.c@culturalinfusion.org.au, nicole.lee@diversityatlas.io, rezza.moieni@diversityatlas.io, mary.legrand@diversityatlas.io

**How to cite this paper:** Chavan, A., Lee, N., Moieni, R., & Legrand, M. (2025). Quantifying Language Disparities: A Distance-Based Predictive Model Using Linguistic Tree Structures. *Open Journal of Social Sciences, 13,* 554-570.
https://doi.org/10.4236/jss.2025.1312040

**Received:** October 13, 2025
**Accepted:** December 28, 2025
**Published:** December 31, 2025







## Abstract

Understanding the historical evolution and divergence of languages requires a quantitative framework for measuring their relationships. Traditional linguistic classifications provide hierarchical structures but lack numerical quantification of language disparity. This project addresses this gap by developing a computational model to quantify linguistic distance within a language tree. This study employs a graph-based model, utilising the Shortest Path Algorithm and Levenshtein Distance to determine the minimum distance between languages, capturing their ancestral and lexical relationships, respectively. This distance is then converted into a relationship score using an exponential decay function, reflecting the observed non-linear pattern of linguistic similarity. The resulting model provides a robust computational tool for understanding language disparity, moving beyond broad classifications to provide numerical insights into how closely languages are related, contributing to the study of language evolution.

## Keywords

Linguistic Evolution, Language Disparity, Shortest Path Algorithms, Exponential Decay, Levenshtein Distance, Computational Linguistics, Quantitative Linguistics, Network Analysis

## 1. Introduction

Modern workplaces are increasingly characterised by multilingualism, with employees communicating across a wide spectrum of languages and dialects. Whether in global corporations, government agencies, or local enterprises, collaboration involves individuals from diverse linguistic and cultural backgrounds. Although

this diversity contributes to the creativity and richness of the workplace, it also introduces communication challenges. As Angouri (2014) notes, multilingualism is not an exception but a defining feature of the contemporary workplace. While English often functions as a common corporate language (Nickerson, 2005; Fredriksson et al., 2006), it rarely eliminates the need to navigate varied linguistic landscapes.

Employees are increasingly expected to collaborate in virtual teams, travel across regions, and complete complex tasks in languages that are often not their first. Language disparity—the measurable difference between languages in terms of structure, vocabulary, and origin—can result in subtle but wide-ranging effects. Employees facing linguistic barriers may experience exclusion from informal networks, reduced participation in decision-making, and limited access to training and career advancement opportunities (Holmes & Stubbe, 2003). Miscommunication caused by language differences can slow workflows, increase cognitive load, and lead to errors, ultimately reducing team productivity and innovation (Dale-Olsen & Finseraas, 2020). Salleh et al. (2021) further highlight that language barriers significantly hinder workplace communication and are strongly correlated with lower productivity and job satisfaction among foreign workers, particularly when discrimination is also present.

The significant impact of language disparity underscores a crucial need for a structured way of measuring it within organisations. To address this, prior research in linguistics and computational language modelling has explored various methods to estimate language similarity. These include scalar distance metrics based on second-language learning difficulty (Chiswick & Miller, 2005), perplexity-based scoring using language models (Bortoletto et al., 2018), and semantic divergence via word embeddings (Asgari & Mofrad, 2016). Other approaches rely on lexical distance measures such as the normalised Levenshtein distance (Brown, 2016), or on grammatical variation modelling within related languages (Szmrecsanyi et al., 2019). While these methods offer useful insights, they are often corpus-dependent, focused on specific language families, or do not reflect the type of lexical and ancestral relationships that affect real workplace communication.

This study proposes a novel, distance-based framework to quantify language disparity in a measurable way. The method integrates two approaches: 1) computing the shortest path between languages within a linguistic family tree and applying an exponential decay function to reflect ancestral distance; and 2) using Levenshtein distance to calculate lexical similarity across core vocabulary words, enabling fine-grained comparisons within language families. By assigning a numerical relationship score to any language pair, this model allows organisations to identify and interpret hidden language divides, offering a data-driven foundation for designing more inclusive workplace communication strategies. To operationalise this framework, we also developed an interactive tool that allows users to input any two languages and instantly retrieve a relationship score, common ancestor, distance path, and lexical similarity, enabling organisations to explore lin-

guistic gaps, support inclusive communication, and design better multilingual collaboration strategies.

### Problem Statement

Languages exhibit varying degrees of relationship through their phonetic, morphological, and syntactic similarities. While established hierarchical language classifications, such as those in David Dalby's *The Linguasphere Register* (Dalby, 2000), provide broad groupings, there is a lack of numerical quantification for how "related" or "distant" languages are.

This gap creates significant challenges, particularly in multilingual workplaces, where organisations struggle to manage the impact of language disparity. Existing computational methods for estimating language similarity often fail to adequately capture the ancestral and lexical relationships for real-world communications. Hence, a structured and computational framework for measuring linguistic distance within a comprehensive language tree is needed.

## 2. Literature Review

### 2.1. Language Classification and Evolution

Languages can be systematically classified using hierarchical models based on their historical development, structural features, and geographic distribution. One of the most comprehensive systems was proposed by David Dalby in *The Linguasphere Register* (Dalby, 2000), which introduced a dual classification framework. This system organises the world's languages into "Sectors" (broader groupings) and "Zones" (more specific clusters), further distinguishing them by:

- **Phylosectors:** Languages that share historical and genetic ancestry, mirroring phylogenetic trees in biology.
- **Geosectors:** Languages that are geographically proximate, acknowledging the influence of regional interaction on linguistic evolution.

Dalby's approach reflects models in evolutionary biology, where species are classified by shared ancestry and environmental adaptation. This structure underscores that languages evolve gradually and are interconnected, rather than existing in isolation.

Moreover, Dalby posited that language variation occurs along a continuum, rather than through abrupt categorical shifts. For example, while English and German share common Germanic ancestry, their distinct phonetic and grammatical structures illustrate a significant degree of divergence. Extending this concept, other more distantly related Indo-European languages like French (from the Romanic zone) and Russian (from the Slavic zone) would exhibit a continuous decrease in shared features yet maintain overlapping traits across the broader linguistic sector. This continuous, gradient-like change is critical for modelling linguistic distance as a measurable, non-binary value, allowing for a more accurate representation of language disparity.

## 2.2. Existing Quantitative Approaches to Linguistic Distance

The study of linguistic distance plays a central role in both theoretical linguistics and applied language science, particularly in understanding language evolution, acquisition difficulty, and structural divergence. While the concept of "distance" between languages was historically qualitative, computational and quantitative methods now enable researchers to assign measurable values to language similarity. This literature review synthesises several key approaches to linguistic distance—ranging from lexical string-based similarity measures to embedding-driven divergence models and hierarchical path-based frameworks—all of which collectively inform the methodology adopted in the present study.

### 2.2.1. Early Foundation: Scalar and Lexical Metrics

Chiswick and Miller (2005) provided one of the earliest systematic efforts to quantify linguistic distance by correlating language difficulty with English proficiency among adult immigrants in the United States and Canada. Their scalar distance model demonstrated that linguistic distance has a measurable impact on real-world language acquisition outcomes.

Specifically, they found that individuals whose native languages were more "distant" from English exhibited significantly lower English proficiency, even after controlling for socioeconomic and educational factors. This distance, while socio-cognitive in effect, was grounded in structural properties such as phonological systems, morphology, and syntactic rules. Their work thus justified the importance of quantifying linguistic proximity beyond intuition, laying foundational groundwork for computational approaches to linguistic measurement.

A key method for quantifying linguistic similarity, and one that provide a fine-grained metric of language relatedness, is Levenshtein Distance (LD). Levenshtein Distance calculates the minimum number of insertions, deletions, or substitutions required to transform one word into another. Gooskens and Heeringa (2004) applied this technique to measure pronunciation differences between dialects of Dutch and Norwegian using phonetic transcriptions. Their findings validated Levenshtein Distance as a reliable quantitative tool, demonstrating that dialects perceived as similar by native speakers also had low Levenshtein Distance, thereby capturing lexical and phonological proximity.

### 2.2.2. Perplexity-Based and Semantic Models

Bortoletto et al. (2018) applied a novel perplexity-based model to measure linguistic distance across five closely related Romance languages: Catalan, Spanish, Galician, European Portuguese, and Brazilian Portuguese. Their method, grounded in probabilistic language modelling, calculates how well a language model trained on one language predicts another. The study revealed that lower perplexity correlated with closer linguistic relationships, aligning well with both geographic and historical expectations. For instance, Brazilian Portuguese exhibited high perplexity relative to Catalan, consistent with both spatial and linguistic divergence. This approach provides an important precedent for quantifying language similarity

through predictability and statistical modelling, complementing the deterministic nature of Levenshtein Distance. Although perplexity requires large corpora and robust language models, it adds a dimension of depth to similarity by indirectly integrating word usage frequency, syntax, and grammar via n-gram statistics.

A more recent and semantically sensitive approach to language comparison is the Word Embedding Language Divergence (WELD) metric, proposed by Asgari and Mofrad (2016). WELD measures the divergence between word vector distributions across languages, typically learned from parallel corpora like Bible translations. Their methodology converts each language into a high-dimensional graph of semantic relationships between words, with the distance between two languages then computed as the statistical divergence between their respective graphs. WELD offers several advantages: it captures both syntactic and semantic similarity, avoids dependency on orthographic similarity, and is extensible to dialect and domain comparison. For example, it could reveal that "mother" and "mum" are semantically identical despite significant lexical differences. While this approach is beyond the scope of the current study due to corpus limitations, it represents a powerful extension of linguistic distance research, particularly in multilingual settings where translation equivalents can be aligned.

#### 2.2.3. Syntactic and Probabilistic Similarity

While lexical comparisons are useful for surface-level similarity, deeper syntactic structures can reveal more nuanced relationships among languages. Szmrecsanyi et al. (2019) introduced the Variation-Based Distance and Similarity (VADIS) framework to assess grammatical and probabilistic similarities between varieties of English spoken around the world. Their study examined how different Englishes—such as British, Indian, Nigerian, and Singaporean English—realise alternating grammatical constructions (e.g., the genitive "John's book" vs. "the book of John", or dative constructions like "give John the book" vs. "give the book to John"). By analysing the probabilistic usage patterns of these structures across dialects, the authors were able to map structural similarity at a grammatical level, rather than purely lexical. They found that even languages that share a common ancestry and vocabulary base may diverge significantly in syntactic preferences, influenced by local substrate languages, colonisation history, and sociolinguistic norms.

This work suggests that true linguistic distance cannot be fully captured by lexical comparison alone. In the context of this study, while the current methodology uses Levenshtein-based lexical distance, future extensions could benefit from incorporating syntactic variation models to provide a multi-dimensional view of language similarity, especially in cases where grammatical convergence or divergence plays a critical role in language evolution

### 2.3. Theoretical Foundations for Distance Decay in Language

The concept of "distance decay"—the principle that similarity between two entities decreases as the distance between them increases—is foundational in many scientific disciplines, including ecology, geography, and linguistics. It provides a

powerful framework for understanding how relatedness between languages changes over time, space, and lineage. In this study, we apply this principle to model how linguistic similarity declines as ancestral or lexical distance grows.

### 2.3.1. Ecological and Cross-Domain Analogies

In Ecology, Nekola and White (1999) formulated the idea that community similarity declines exponentially with increasing geographic distance. For example, two neighbouring forest patches will share more plant and animal species than two patches separated by larger distances (see Figure 1). This decline is typically modelled using exponential decay functions, which capture two core observations:

1) sharp similarity drops at close distances (due to local differentiation) and

2) slower decay across larger distances (as differences accumulate more gradually).

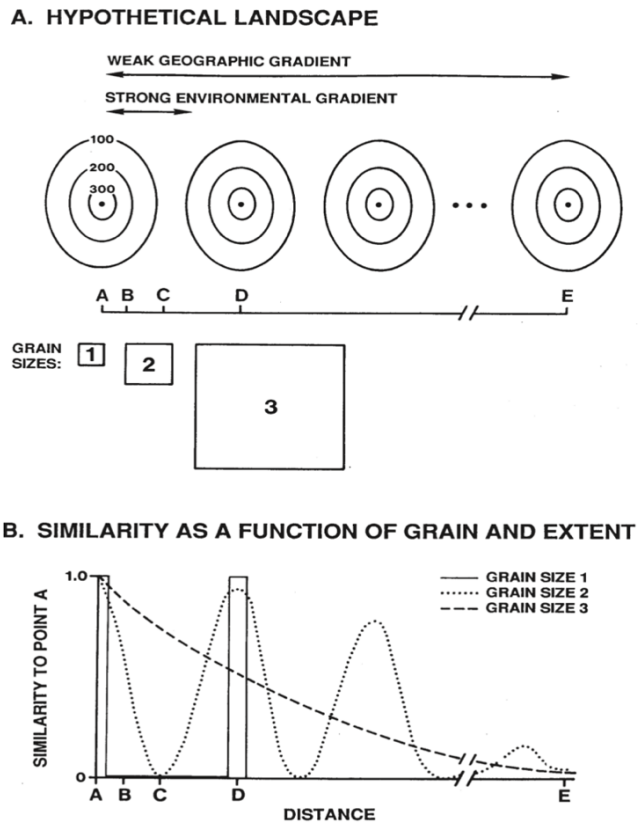



Figure 1. Graph of similarity as a function of grain and extent from Nekola and White's study. It quantifies how biological communities become less similar as spatial distance increases.

This non-linear behaviour was later confirmed by Dias et al. (2021), who modelled species distribution in river networks using Bayesian approaches. Their work emphasised that distance decay is not merely spatial but also relational, accounting for how connected systems (like rivers—or languages) diverge based on

branching paths. This principle provides a compelling analogy for linguistic evolution: just as ecosystems become less alike over space, languages diverge over genealogical and structural distance.

### 2.3.2. Linguistic Evidence for Non-Linear Decay

The phenomenon of non-linear decay in similarity is also empirically supported within linguistics. Lieberman et al. (2007) provided a quantitative analysis of historical linguistic evolution by examining how English verb forms changed over 1200 years. Their model demonstrated that irregular verbs tend to "regularise" over time at rates inversely related to their frequency, specifically finding that the half-life of an irregular verb is proportional to the square root of its frequency (see Figure 2). This supports the idea that linguistic rules are subject to exponential decay dynamics, where less frequent exceptions are more likely to be assimilated into regular paradigms over time.

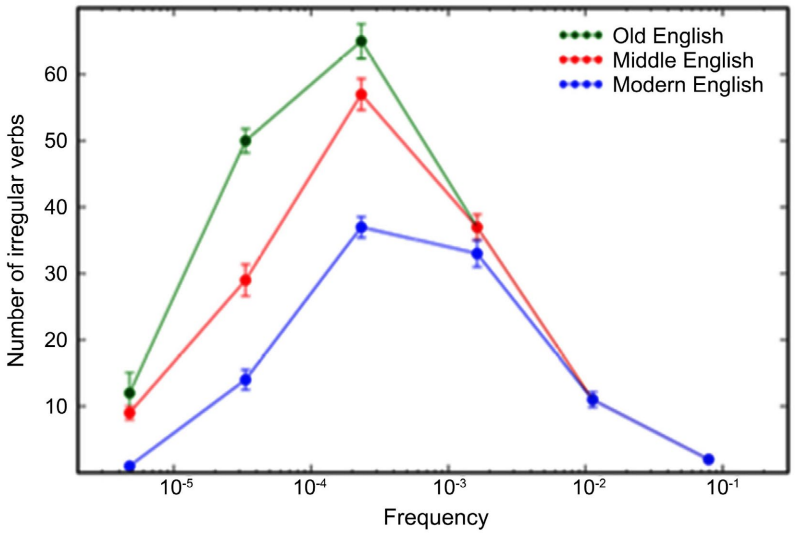



**Figure 2.** Exponential graph for verb irregularity from Lieberman et al. study. It shows that irregular verbs disappear at a rate proportional to their frequency, following an exponential decay model, $N(t) = N0e^{-kt}$.

Similarly, Heeringa (2004) applied Levenshtein Distance to Dutch dialects and observed that dialects physically or historically closer had lower phonetic distances. In cross-linguistic studies, languages that diverged more recently (e.g., Hindi and Urdu) retain more lexical overlap than those that split earlier (e.g., Hindi and English). This decline in similarity is not linear, making exponential decay a more accurate model. Further, Zymet (2014) and Kimper (2011) explored how phonological processes weaken with increased distance, demonstrating that the likelihood of assimilation between two phonemes declines steeply when more syllables intervene, with exponential decay better fitting empirical data than linear models. More recently, a study published in Nature Ecology & Evolution (Bromham et al., 2024) found that phoneme inventories across island languages become increasingly dissimilar as geographical distance increases, especially when separated by water

boundaries, supporting the idea that linguistic divergence behaves like ecological speciation, as in strongest close to the origin and more gradual over greater distances.

## 3. Model Development

This study supports a distance-based approach to quantify language disparity by combining hierarchical linguistic relationships with lexical similarity metrics. The methodology involves four major components:

1) constructing a language family tree graph,
2) computing ancestral distance using the shortest path algorithm,
3) comparing lexical similarity using Levenstein distance,
4) modelling similarity through an exponential decay function.

This approach provides a numeric score for the linguistic distance between any two languages.

### 3.1. Language Family Tree Construction

To establish the hierarchical relationships between languages, a comprehensive, static language tress dataset from Diversity Atas is used. This dataset represents languages as nodes and their genealogical or classificatory relationships as edges, mirroring the principle of established hierarchical classification systems such as David Dalby's The Linguasphere Register (Dalby, 2000). This structured representation allows for the mapping of language evolution and the identification of common linguistic origins, transforming the family trees into a quantifiable graph suitable for computational analysis.

### 3.2. Ancestral Distance Calculation: Shortest Path Algorithm

The ancestral distance (d) between any two languages within the family tree is computed using a Shortest Path Algorithm. This algorithm is chosen as the best method for measuring linguistic distance because it effectively captures the hierarchical structure of language evolution, ensuring that ancestral relationships and linguistic divergence are reflected.

By computing the minimum number of steps or edges required to traverse from one language node to another via their nearest common ancestor, it mirrors genetic divergence in evolutionary linguistics. This ensures distance is measured based on shared ancestry and structural distance, rather than random string-matching or geographic-only models.

The application of the Shortest Path Algorithm is adapted based on the ancestral relationship, with two distinct computational approaches used depending on the degree of relatedness:

- **For languages belonging to different major ancestral backgrounds**:

The standard Shortest Path Algorithm identifies the minimal path between the two language nodes, tracing back to their lowest common ancestor and then descending to the largest language. The distance (d) is defined as the total number

of edges along this path. Figure 3 below presents an example of tracking the path between English and Zulu.

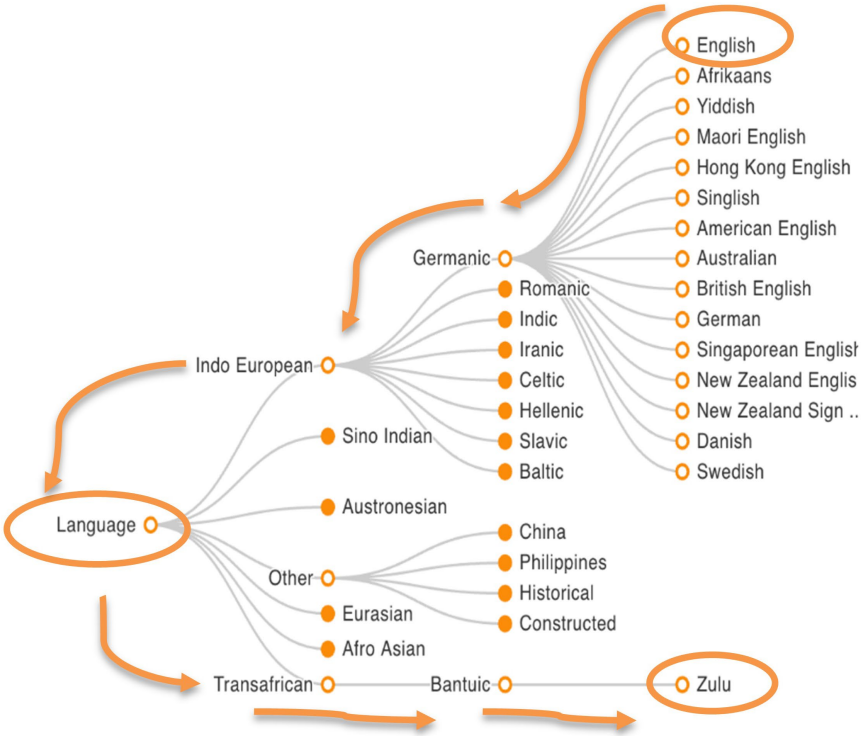



Figure 3. Example of tracking shortest path between English and Zulu. The distance between the two languages is six steps.

- **For languages within the same family sharing a very close common ancestor (e.g., dialects or closely related languages):**

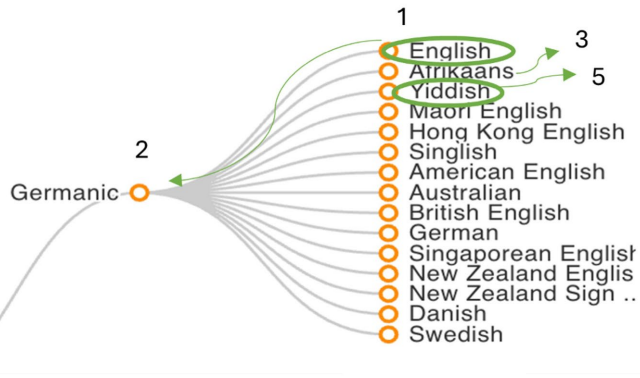



Figure 4. Example of tracking shortest path between English and Yiddish. The distance between the two languages is five steps.

A Depth-First Search (DFS) approach is implemented instead. DFS is used because closely related languages often have multiple intermediate variants, dialects, or sibling branches within the same family tree. The DFS method systematically traverses from one language to another, visiting all relevant sibling nodes and in-

termediate branches to construct a complete path through the common ancestor. This ensures the computed distance captures the full complexity of relationships within a tightly connected language family, including all intermediate divergences that would be overlooked by a simple shortest part calculation. Figure 4 below presents an example of tracking the path between English and Yiddish sitting in the same Germanic family.

### 3.3. Lexical Similarity: Levenshtein Distance

Lexical similarity is a measure of how similar the words are in two languages based on their spelling or structure, as related languages often share word roots and spelling patterns. On top of the tree-based ancestral modelling, Levenshtein Distance (LD) is used to arrange languages within the languages tree that *share the same common ancestor* into a logical order based on their word-level similarity.

By comparing a set of four core vocabulary words—water, house, sun, and hello—it calculates the minimum number of single-character edits required to translate one word into another. These words were selected to represent fundamental human concepts: natural elements (water, sun), built environment (house), and basic social interaction (hello), following principles similar to the Swadesh list approach, which prioritises basic, culturally stable terms that are present across most languages and less susceptible to borrowing or rapid semantic change. The edits include:

- Insertion: adding a character
- Deletion: removing a character
- Substitution: changing one character to another

The process for determining lexical similarity involves two key steps:

#### 3.3.1. Data Generation

Common vocabulary words in various languages were obtained using Claude AI, which was prompted to provide translations of a standardised list of the four core vocabulary items across target languages within the language tree dataset. These translations we generated through structures prompt requesting word-for-word equivalents in each target language. All translations were normalised to lowercase to ensure consistency in subsequent distance calculations. This provided the foundational dataset of equivalent words across different languages for comparative analysis. Table 1 shows an extract of the dataset showing vocabularies from four Germanic languages within the Indo-European family.

Table 1. Extract from the dataset containing translations of common words: water, house, sun, and hello in Aachenisch, Aboriginal English, Achterhoeks, and Afrikaans.

| Languages | Translations | | | |
|---|---|---|---|---|
| | water | house | sun | hello |
| Aachenisch | wasser | huus | sonn | hallo |
| Aboriginal English | water | house | sun | yaama |
| Achterhoeks | water | hoes | zunne | moi |
| Afrikaans | water | huis | son | hallo |

### 3.3.2. Levenshtein Distance Calculation

For each pair of corresponding words across two languages ($w_1$, $w_2$), the raw LD is calculated. The distance is then normalised by the maximum length of the two words.

The formulas applied are:

- Distance (D) = Levenshtein Distance ($w_1$, $w_2$)
- Max Length (L) = max(len($w_1$), len($w_2$))
- Normalised Distance = D/L
- Similarity Score = 1 – Normalised Distance

For a set of word pairs between two languages, the average similarity is then computed as the mean of the individual word pair similarity scores. This average score quantifies the overall lexical overlap between the two languages.

Table 2 shows examples of calculating LD between Afrikaans and Aachenisch. For example, transforming from "water" to "wasser" requires two edits: (1) inserting an "s" after "a" in "water" (resulting in "waster"), and (2) substituting "t" in "waster" with "s" (resulting in "wasser"). With the maximum length being six characters from "wasser", the similarity score for these two terms is then 1 – 2/6 = 0.67. Average similarity is the mean of (0.67 + 0.75 + 0.75 + 1.00) = 0.7925. Similarly, the model calculates for Afrikaans verses all other languages.

Table 2. Example of calculating LD between Afrikaans and Aachenisch.

| Afrikaans | Aachenisch | Distance (D) | Max Length (L) | Similarity Score |
|---|---|---|---|---|
| water | wasser | 2 | 6 | 1 – 2/6 = 0.67 |
| huis | huus | 1 | 4 | 1 – 1/4 = 0.75 |
| son | sonn | 1 | 4 | 1 – 1/4 = 0.75 |
| hallo | hallo | 0 | 5 | 1 – 0/5 = 1.00 |

Table 3 shows another example of calculating LD between Afrikaans and Aboriginal English. Average Similarity is the mean of (1.00 + 0.40 + 0.67 + 0.00) = 0.5175.

Table 3. Example of calculating LD between Afrikaans and Aboriginal English.

| Afrikaans | Aboriginal English | Distance (D) | Max Length (L) | Similarity Score |
|---|---|---|---|---|
| water | water | 0 | 5 | 1 – 0/5 = 1.00 |
| huis | house | 3 | 5 | 1 – 3/5 = 0.40 |
| son | sun | 1 | 3 | 1 – 1/3 = 0.67 |
| hallo | yaama | 5 | 5 | 1 – 5/5 = 0.00 |

These two examples demonstrate how LD quantifies similarity between languages in the same language family. From the above examples, Afrikaans and Aachenisch exhibit a higher lexical similarity of 0.7925, indicating their closer grouping, whereas Afrikaans and Aboriginal English show a lower similarity of

0.5175, resulting in their placement further apart.

While LD does not inherently account for grammatical structure or historical lineage, it offers a practical and interpretable measure of word-level difference, which is relevant when evaluating language disparity in workplace communication. By complementing the tree-based ancestral distance modelling with LD, it allows both lexical form and historical context to inform the final relationship score.

### 3.4. Relationship Score Modelling: Exponential Decay Function

The ancestral distance (d) derived from the shortest path algorithm is transformed into a quantifiable relationship score using an exponential decay function. This function is chosen to reflect the non-linear pattern of linguistic similarity, where similarity decreases rapidly at short distances and more gradually over longer distances. This behaviour aligns with studies on linguistic variation, including those employing Levenshtein distance, which show a steep decline in similarity at short distances and a more gradual decline at larger distances.

The relationship score is calculated using the exponential decay formular below:

$$\text{Similarity Score} = e^{-d} \tag{1}$$

where:

- d = the shortest path distance distance between two language nodes on the language tree
- e = Euler's number (approximately 2.718)

By adopting exponential decay, the model ensures the relationship score starts high for closely related languages (small d) and declines quickly as linguistic divergence increase. This approach aligns with ecological models of species divergence, phonological studies demonstrating how linguistic processes weaken with increasing distance and observed language tree evolution in linguistics.

To categorise language relationships in a meaningful way, threshold values are applied to exponential decay to distinguish between "closely related" or "distantly related" languages pairs. The model uses the exponential decay function (Similarity Score = $e^{-d}$ where d represents the tree distance between languages. Two threshold approaches are offered: a strict threshold using $e^{-1} \geq 0.368$, which classifies only language pairs one step apart (d = 1) as closely related, and a more relaxed, and more relaxed threshold $e^{-2} \geq 0.135$, which extends this classification to language pairs up to two steps apart ($d \leq 2$). These cutoff values were chosen because they correspond to natural division in language family structures. A distance of d=1 represents languages like Spanish and Portuguese, which typically share substantial mutual intelligibility. A distance of d = 2 captures the next tier of relationship, with broader subfamilies like Germanic or Slavic groups, which share common roots but have diverged significantly. This threshold-based approach provides flexibility for different analytical needs while remaining consistent with the observed non-linear pattern of linguistic divergence.

## 4. Discussion

### 4.1. Model Advantages

This integrated approach uses the shortest path algorithm to measure hierarchical distance and applies Levenshtein Distance (LD) to compute a nuanced relationship score. This methodology provides a robust framework by realistically aligning with established species divergence models, observed language tree evolution, and lexical similarity.

The exponential decay function specifically addresses the limitations of alternative decay models when mimicking real-world language evolution:

- Linear decay assumes a constant rate of change, which fails to capture the fast initial divergence observed in language trees.
- Logarithmic decay suggests a slow initial change followed by faster divergence, which contradicts observed patterns in language trees.
- Polynomial decay can often oversimplify or misrepresent the non-linear nature, potentially suggesting languages become completely distinct too early or plateauing unnaturally.

While the proposed model shares conceptual foundations with existing approaches—such as the lexical string comparison used by Gooskens and Heeringa (2004) and the hierarchical classification principles of Dalby (2000)—it offers distinct advantages through its integration of multiple dimensions. Unlike perplexity-based models (Bortoletto et al., 2018) which require large parallel corpora and robust language models, our approach operates on a minimal vocabulary set and a static tree structure, making it computationally efficient and scalable across thousands of languages. Compared to semantic embedding approaches like WELD (Asgari & Mofrad, 2016), which capture deeper semantic relationships but depend on extensive parallel texts, our method provides immediate, interpretable distance metrics based on both ancestral and lexical relationships. The key innovation lies in combining tree-based path distance with Levenshtein similarity through a unified exponential decay framework, enabling the model to capture both historical divergence and present-day lexical proximity in a single, interpretable score suitable for organisational decision-making.

For practical application in multilingual workplace settings, the numerical relationship scores can be interpreted as indicators of communication difficulty and training resource allocation. For example, a relationship score above 0.368 (threshold for $d = 1$) suggests that employees speaking these language pairs may share substantial mutual intelligibility and require less intensive language support, while scores below 0.135 (threshold for $d = 2$) indicate significant linguistic distance requiring more comprehensive translation services, extended training periods, or specialised communication strategies. Managers can use these scores to prioritise language pairing in team assignments, estimate cross-language communication costs, and design targeted language training programs based on the degree of linguistic disparity between employee language backgrounds.

### 4.2. Interactive Tool for Practical Application

To deploy this framework into a practical application, an interactive tool has been developed under Diversity Atlas domain. This tool allows user to input any two or more (up to four) languages and instantly calculate their relationship score. It also identifies their common ancestor, visualises the distance path within the language tree, and displays their lexical similarity based on Levenshtein Distance. Figure 5 below shows screenshot of the web application.

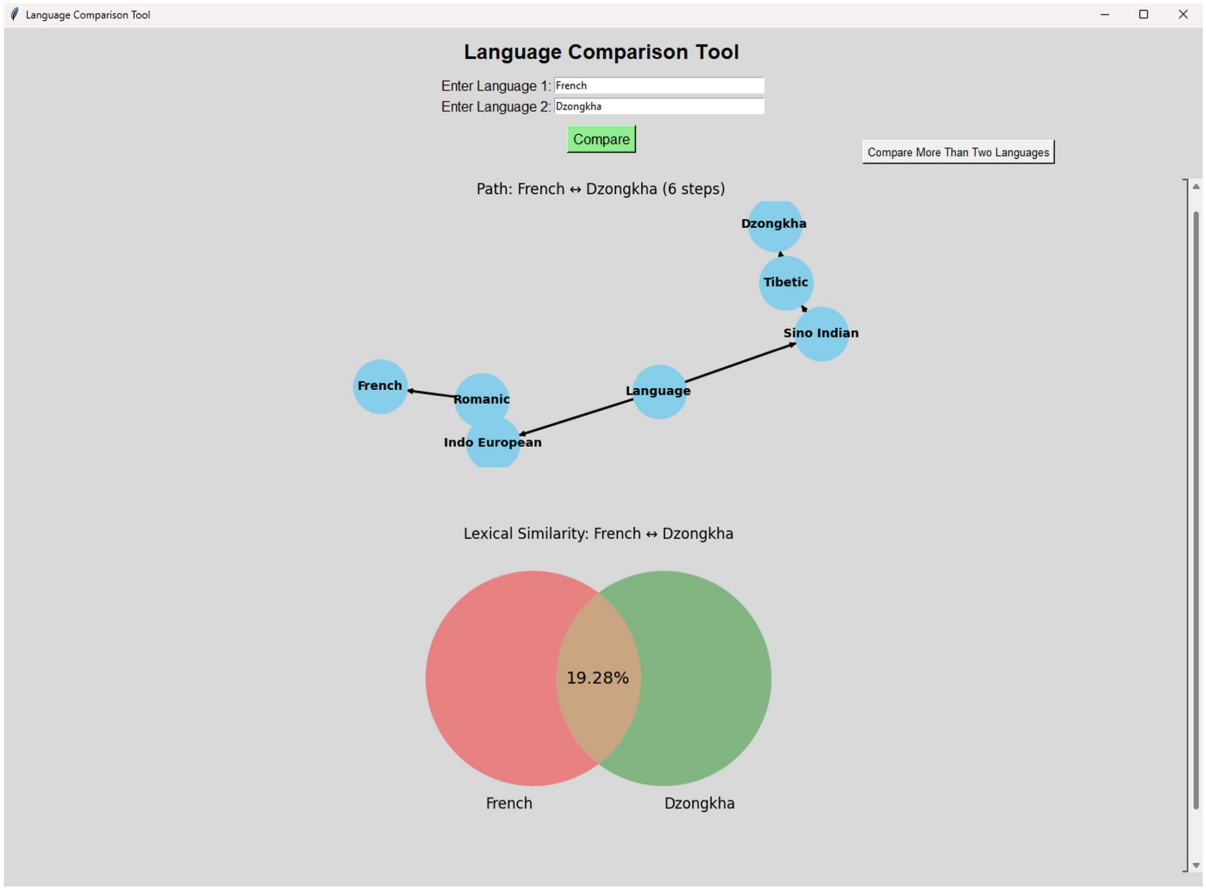



Figure 5. Screenshot of the web application.

The tool integrates the linguistic tree with word-level analysis and presents results visually through path graphs and Venn diagrams, making the concept of language disparity more tangible and actionable for organisations seeking to design more inclusive multilingual communication strategies.

### 4.3. Limitations and Future Work

This model relies on Large Language Model (LLM) generated translations for lexical similarity, which might not always be accurate, particularly for rare languages or dialects. Besides, it does not include sign languages, which represent a distinct category of human communication. Their unique characteristics would need a different approach to quantify distance.

Looking ahead, there are a couple of refinements that can be done to enhance the model and expand its utility to cove more languages. Future work could focus on validating LLM-generated translation with subject matter experts for rare languages. This would reduce potential inaccuracies and strengthen the reliability of the lexical similarity component. We may also extend the model to include sign languages by developing methods to capture their unique visual and spatial elements, such as handshapes and facial expressions, which differ from spoken languages. Finally, the current model uses a static language tree provided by Diversity Atlas, for which future work could explore integrating dynamic language trees that evolve with new languages discovered, offering a more scalable and adaptable framework.

## 5. Conclusion

This paper addresses the need for a quantifiable approach to measuring linguistic disparity, by introducing an integrated framework that combines hierarchical linguistic relationships from established language family trees with lexical similarity metrics derived from Levenshtein Distance. We also apply the shortest path algorithm to compute ancestral distance and an exponential decay function to translate this distance into a single score. This method aligns with observed patterns of language evolution and species divergence in Ecology.

By developing the model, it offers a significant contribution to computational linguistics, providing a practical way to quantify the interconnection and divergence of languages. It could help optimise multilingual communication in global workplaces to inform inclusive strategies and enrich our understanding of the dynamics of linguistic changes.

While the current model establishes a strong foundation, future work will focus on refining lexical data sourcing and exploring ways to include sign languages for a more holistic view of linguistic distance.

## Acknowledgements

This research was supported by Diversity Atlas and their provision of data has been instrumental in shaping the findings of this study. We thank the entire Diversity Atlas team for their support, although they may not agree with all the interpretations/conclusions of the results.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Angouri, J. (2014). Multilingualism in the Workplace: Language Practices in Multilingual Contexts. *Multilingua, 33,* 1-9. https://doi.org/10.1515/multi-2014-0001

Asgari, E., & Mofrad, M. R. K. (2016). Comparing Fifty Natural Languages and Twelve Genetic Languages Using Word Embedding Language Divergence (WELD) as a Quan-

titative Measure of Language Distance. In *Proceedings of the Workshop on Multilingual and Cross-Lingual Methods in NLP* (pp. 65-74). Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-1208

Bortoletto, G., Manninen, L., McKenzie, E., & Raatikainen, O. (2018). *Measuring Language Distance Using Perplexity*. Natural Language Engineering. https://raaoo.github.io/assets/documents/linguistic_distance.pdf

Bromham, L., Yaxley, K. J., & Cardillo, M. (2024). Islands Are Engines of Language Diversity. *Nature Ecology & Evolution, 8,* 1991-2002. https://doi.org/10.1038/s41559-024-02488-4

Brown, S. (2016). Two Measures of Linguistic Distance. *Cultural Anthropology and Ethnosemiotics, 2,* 2-12. https://www.researchgate.net/publication/308609789_Two_measures_of_linguistic_distance

Chiswick, B. R., & Miller, P. W. (2005). Linguistic Distance: A Quantitative Measure of the Distance between English and Other Languages. *Journal of Multilingual and Multicultural Development, 26,* 1-11. https://doi.org/10.1080/14790710508668395

Dalby, D. (2000). *The Linguasphere Register of the World's Languages and Speech Communities*. Ross. https://hdl.handle.net/11858/00-001M-0000-0012-7807-B

Dale-Olsen, H., & Finseraas, H. (2020). Linguistic Diversity and Workplace Productivity. *Labour Economics, 64,* Article ID: 101813. https://doi.org/10.1016/j.labeco.2020.101813

Dias, F. S., Betancourt, M., Rodríguez-González, P. M., & Borda-de-Água, L. (2021). Analysing the Distance Decay of Community Similarity in River Networks Using Bayesian Methods. *Scientific Reports, 11,* Article No. 21660. https://doi.org/10.1038/s41598-021-01149-x

Fredriksson, R., Barner-Rasmussen, W., & Piekkari, R. (2006). The Multinational Corporation as a Multilingual Organization: The Notion of a Common Corporate Language. *Corporate Communications: An International Journal, 11,* 406-423. https://doi.org/10.1108/13563280610713879

Gooskens, C., & Heeringa, W. (2004). Perceptive Evaluation of Levenshtein Dialect Distance Measurements Using Norwegian Dialect Data. *Language Variation and Change, 16,* 189-207. https://doi.org/10.1017/s0954394504163023

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. https://pure.rug.nl/ws/portalfiles/portal/9800656/thesis.pdf

Holmes, J., & Stubbe, M. (2003). *Power and Politeness in the Workplace: A Sociolinguistic Analysis of Talk at Work*. Routledge.

Kimper, W. A. (2011). Locality and Globality in Phonological Variation. *Natural Language & Linguistic Theory, 29,* 423-465. https://doi.org/10.1007/s11049-011-9129-1

Lieberman, E., Michel, J., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the Evolutionary Dynamics of Language. *Nature, 449,* 713-716. https://doi.org/10.1038/nature06137

Nekola, J. C., & White, P. S. (1999). The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography, 26,* 867-878. https://doi.org/10.1046/j.1365-2699.1999.00305.x

Nickerson, C. (2005). English as a Lingua Franca in International Business Contexts. *English for Specific Purposes, 24,* 367-380. https://doi.org/10.1016/j.esp.2005.02.001

Salleh, M. M., Mohi, Z., Nordin, N., Mohamad, N. A., & Razali, N. A. S. (2021). The Impact of Language Barriers and Discrimination Issues on Work Productivity of Foreign Workers. *International Journal of Academic Research in Business and Social Sciences, 11,* 42-

52. https://doi.org/10.6007/ijarbss/v11-i16/11215

Szmrecsanyi, B., Grafmiller, J., & Rosseel, L. (2019). Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes. *Frontiers in Artificial Intelligence, 2,* Article No. 23. https://doi.org/10.3389/frai.2019.00023

Zymet, J. A. (2014). *Distance-Based Decay in Long-Distance Phonological Processes: A Probabilistic Model for Malagasy, Latin, English, and Hungarian*. University of California. https://escholarship.org/uc/item/20n7n3gj

## Appendix. GitHub Link

The Language tree datasets and the codes for computing distances are stored in the Cultural Infusion GitHub repository. Access is granted upon request https://github.com/CulturalInfusion/language-distance