Chinese “dialects” and European “languages”: a comparison of lexico-phonetic and syntactic distances

Tang, C.; Heuven, V.J.J.P. van; Heeringa, W.; Gooskens, C.

languages

MDPI

Article

Chinese "Dialects" and European "Languages": A Comparison of Lexico-Phonetic and Syntactic Distances

Chaoju Tang [1], Vincent J. van Heuven [2,3,4,*], Wilbert Heeringa [4] and Charlotte Gooskens [5]

[1] School of Foreign Languages, University of Electronic Science and Technology, Chengdu 611731, China; tangchaoju1968@uestc.edu.cn
[2] Leiden University Centre for Linguistics, 9500 RA Leiden, The Netherlands
[3] Doctoral School of Multilingualism, University of Pannonia, H-8210 Veszprém, Hungary
[4] Fryske Akademy, 8911 DX Leeuwarden, The Netherlands; wheeringa@fryske-akademy.nl
[5] Center for Languages and Cognition, University of Groningen, 9700 AS Groningen, The Netherlands; c.s.gooskens@rug.nl
* Correspondence: v.j.j.p.van.heuven@hum.leidenuniv.nl

Abstract: In this article, we tested some specific claims made in the literature on relative distances among European languages and among Chinese dialects, suggesting that some language varieties within the Sinitic family traditionally called dialects are, in fact, more linguistically distant from one another than some European varieties that are traditionally called languages. More generally, we examined whether distances among varieties within and across European language families were larger than those within and across Sinitic language varieties. To this end, we computed lexico-phonetic as well as syntactic distance measures for comparable language materials in six Germanic, five Romance and six Slavic languages, as well as for six Mandarin and nine non-Mandarin ('southern') Chinese varieties. Lexico-phonetic distances were expressed as the length-normalized MPI-weighted Levenshtein distances computed on the 100 most frequently used nouns in the 32 language varieties. Syntactic distance was implemented as the (complement of) the Pearson correlation coefficient found for the PoS trigram frequencies established for a parallel corpus of the same four texts translated into each of the 32 languages. The lexico-phonetic distances proved to be relatively large and of approximately equal magnitude in the Germanic, Slavic and non-Mandarin Chinese language varieties. However, the lexico-phonetic distances among the Romance and Mandarin languages were considerably smaller, but of similar magnitude. Cantonese (Guangzhou dialect) was lexico-phonetically as distant from Standard Mandarin (Beijing dialect) as European language pairs such as Portuguese–Italian, Portuguese–Romanian and Dutch–German. Syntactically, however, the differences among the Sinitic varieties were about ten times smaller than the differences among the European languages, both within and across the families—which provides some justification for the Chinese tradition of calling the Sinitic varieties dialects of the same language.

Keywords: affinity trees; Chinese dialects; European languages; Levenshtein distance; lexico-phonetic distance; multi-dimensional scaling (MDS); Pointwise Mutual Information (PMI); PoS-tag trigram frequency; syntactic distance

Academic Editors: Alfred Lameli, Simonetta Montemagni and John Nerbonne

Received: 3 September 2024
Revised: 14 May 2025
Accepted: 22 May 2025
Published: 29 May 2025

Citation: Tang, C., van Heuven, V. J., Heeringa, W., & Gooskens, C. (2025). Chinese "Dialects" and European "Languages": A Comparison of Lexico-Phonetic and Syntactic Distances. *Languages*, *10*(6), 127. https://doi.org/10.3390/languages10060127

1. Introduction

It has often been claimed, or observed anecdotally, that the differences among European languages may well be smaller than the differences among language varieties in China that are referred to as dialects by Chinese linguists.

Consider, for instance, the following passage, taken from Wardhaugh (2008, p. 32):

> Speakers of Cantonese and Mandarin will tell you that they speak the same language. However, if one speaker knows only Cantonese and the other only Mandarin, they will not be able to converse with each other: they actually speak different languages, certainly as different as German and Dutch and even Portuguese and Italian. If the speakers are literate, however, they will be able to communicate with each other through a shared writing system. They will almost certainly insist that they speak different dialects of Chinese, not different languages.

This passage echoes an earlier version of the claim made by C. Li and Thompson (1981, p. 2):

> It is traditional to speak of the different varieties of Chinese as "dialects", even though they may be different from one another to the point of being mutually unintelligible. It is often pointed out, for example, that Cantonese and Mandarin differ from each other roughly as the Romance "languages" Portuguese and Rumanian do. On the one hand, because Portuguese and Rumanian are spoken in different countries, they are referred to as different "languages". On the other hand, because Cantonese and Mandarin are spoken in the same country, they are called different "dialects".

Various reasons have been given to explain the tradition in Chinese linguistics to consider the varieties mentioned above dialects. These are alluded to in the Wardhaugh passage above. The first argument is partly cultural and partly political. To most Chinese citizens, wherever they live in the country, all of China is one large country with one overarching culture and one language that unifies the entire nation. In such a cultural tradition, it makes sense to consider all varieties dialects of one language, i.e., Chinese. The other argument is rather more linguistic in nature. Since Chinese is primarily written with logograms rather than spelled alphabetically, the same set of logograms (here, characters) can be used for any different variety of the language, since the characters correspond directly with meanings and have no direct relationship with the phonology of the language.

In spite of the rather insistent and repeated claims about the relative magnitude of the structural differences among European languages and Chinese dialects, there are no published data on the linguistic distances within and between the languages and language families at issue, which would enable a direct evaluation of the claims. It is true that anthropological linguists have come up with lexical comparisons of a large sample taken from the languages of the world, and have used these lexical statistics to substantiate rather more intuitive divisions of the world's languages into families. However, these comparisons are based on a small set of words taken from a restricted domain, i.e., inheritance words (such as family relationships and natural phenomena, Swadesh, 1952), rather than on frequency of use. Moreover, these studies (e.g., Dyen et al., 1992; Gray & Atkinson, 2003; Dunn et al., 2005) quantify the lexical overlap between languages but not the opacity of cognates (phonological distance) or differences in word and sentence structure. A partial solution to these problems is provided by the Automatic Similarity Judgment Program (ASJP, Wichmann et al., 2022), which computes a phonetic similarity measure for all pairs of some 1500 languages based on the transcribed words for 40 concepts. The concepts are a subset drawn from the original Swadesh-100 list. However, the phonetic transcriptions recognize only 7 different vowel qualities, while the 34 different consonant categories often conflate voiced and voiceless counterparts. Stress and tone are not transcribed, nor are there any diacritics (with the exception of nasalization and glottalization marks). Although the ASJP is generally successful in predicting genealogical relationships among the languages in the sample (Pompei et al., 2011), we argue that a better estimate of linguistic similarity can be obtained by using a more refined phonetic distance measure and that differences in syntactic structure should also be included.[1]

There are more sophisticated and diversified quantifications of lexical, phonological and even syntactic distances among European languages (e.g., Swarte, 2016 for Germanic languages; Golubović, 2016 for Slavic languages; Heeringa et al., 2023 for European languages), and lexical and phonological distances among Chinese dialects (e.g., Cheng, 1997; Tang, 2009; Tang & van Heuven, 2015), but the quantifications of the European languages are based on very different materials than those analyzed for the Chinese varieties (Gooskens & van Heuven, 2021). Moreover, although syntactic differences have been studied in a comparison of two word order features (ADP–Noun vs. Noun–ADP and Verb–Object vs. Object–Verb) in Austronesian, Indo-Germanic, Uto-Aztecan and Bantu languages (Dunn et al., 2005, 2011), no information is available on syntactic distance among Chinese varieties.

The purpose of the present paper is not to challenge the idea that mutual intelligibility might be a useful criterion to separate language from dialect (as is implicit in the above passages), nor do we wish to invalidate the use of the term "dialect" for the Chinese varieties mentioned. Instead, we aim to check whether indeed Chinese language varieties, specifically six Mandarin and nine non-Mandarin dialects, including Beijing (or Standard) Mandarin and Cantonese (Guangzhounese), are structurally as different from one another as European languages are, such as German and Dutch, or Portuguese and Italian, or even Portuguese and Romanian—as was claimed in the citations above. We will try to accomplish this by applying the methods of data collection and analysis that we developed earlier in the Micrela project (van Heuven et al., 2015; Gooskens & van Heuven, 2017; Gooskens et al., 2018; Gooskens, 2024) for a selection of European languages (five Germanic, five Romance, six Slavic) to new materials collected for a selection of Sinitic languages/dialects (six Mandarin, nine non-Mandarin).

Importantly, the materials collected for the European and Sinitic languages are equivalent, and can be conceived of as a parallel translation corpus. Linguistic distances among the languages will be estimated separately along two principally different dimensions. The first dimension is probably the most important predictor of successful recognition of (isolated) words in a non-native but closely related language. This dimension is a compounding of the degree of lexical overlap (percentage of cognates shared between two varieties) and the phonological difference between members of cognate pairs. When the same concept is expressed in two languages by non-cognate words, any phonological similarity between these words will be accidental.[2] When the same concept is covered by cognates, the phonological difference will be reduced—depending on the number of sound changes that have taken place in the past since the varieties under comparison split apart. This lexicophonetic dimension is implemented by applying a specific variant of the Levenshtein string edit distance computation (weighting segment distances by Pointwise Mutual Information, PMI, Wieling et al., 2012; Jäger, 2013; see Section 2.2.1 for details).

The second dimension does not so much address the recognition of isolated words, but the reconstruction of the speaker's message from the words recognized in their specific early-to-late order, i.e., the effects of differences in word order on the ease of understanding a related non-native language. Rather than adopting the syntactic parameter approach (determining which specific syntactic features are used in a particular language—as in Longobardi, 2005; Dunn et al., 2005, 2011; Wichmann & Saunders, 2007; Polyakov et al., 2009; Gray et al., 2010; Ceolin et al., 2020, 2021), we opt for a more comprehensive study of surface-syntactic differences between pairs of languages (as proposed by Nerbonne & Wiersma, 2006), counting and subsequently comparing the frequencies of all trigrams of lexical categories (also known as Part of Speech or PoS tags) found in the parallel corpus of texts.

2. Method

The linguistic distance measurements discussed in this paper are based on materials originally collected to assess mutual intelligibility among closely related European languages (Golubović, 2016; Gooskens & van Heuven, 2017, 2020; Gooskens et al., 2018; Swarte, 2016). For the current study, we expanded this material by gathering comparable data from various Chinese language varieties. In Section 2.1, we describe the data set, and in Section 2.2, we outline the methods we used to measure linguistic distances.

2.1. Data Set

2.1.1. Language Varieties

The earlier study included all official national languages from the three largest language families (in terms of number of speakers) in the EU member states: five Germanic languages (Danish, Dutch, English, German, and Swedish), five Romance languages (French, Italian, Portuguese, Romanian, and Spanish), and six Slavic languages (Bulgarian, Croatian, Czech, Polish, Slovak, and Slovene). For the present study, we added West Frisian to the Germanic group as an example of a minority language that is officially recognized by the EU.[3] In cases where a language is official in multiple member states, the variety from the country with the largest speaker population was selected. For instance, although German is an official language in both Austria and Germany, only Northern Standard German from Germany, where the majority of speakers reside, was included.

We included fifteen Chinese varieties, i.e., six from the Mandarin group: Beijing, Jinan, Xi'an, Taiyuan (Northern), Chengdu, and Wuhan (South-Western), as well as nine from the Southern group: Suzhou, Wenzhou (Wu family), Nanchang (Gan family), Meixian (Hakka family), Xiamen, Fuzhou, Chaozhou (Min family), Changsha (Xiang family), and Guangzhou, also known as Cantonese (Yue family). This is a subset of Cheng (1997), which includes at least one dialect spoken in a large city in each of the eight subdivisions of the Sinitic language group (indicated in parentheses in the preceding sentence) customarily distinguished in Chinese dialectology (e.g., C. Li & Thompson, 1981, p. 3; Tang & van Heuven, 2009, p. 712). The 15 dialects are identified by the name of the largest city they are spoken in. These cities have between 3.7 (Nanchang) and 21.5 million (Beijing) inhabitants, with the exception of Meixian (0.6 million).

2.1.2. Materials

To ensure linguistic distance measurements were comparable across language families, we needed to compile equivalent materials for all language varieties. To measure lexico-phonetic distances, we created a list of the 100 most frequently used nouns from the British National Corpus (BNC Consortium, 2007). The list was slightly modified to remove word pairs with overlapping meanings (e.g., in many Slavic languages, the most common translation for both 'work' and 'job' is the same word). When a word was excluded, it was replaced by the next most frequent word from the list. The 100 English words and their Mandarin equivalents are included in Appendix A. Native speakers of each language variety, with some background in translation, translated the English word list into the 16 other European Languages and 15 Chinese language varieties. Each list was initially translated by one native speaker and subsequently checked by at least two others to ensure consensus on the final translation. Finally, broad phonetic transcriptions were made of the 32 word lists by means of pronunciation dictionaries (X. Zhang, 2009 for Chaozhou; R. Li, 1993–1999 for all other varieties) and again checked and corrected by native speakers with a background in phonetics.

To measure syntactic distances, we selected four short English texts from a set of exercises developed by the University of Cambridge to prepare students for the Preliminary

English Test (PET). These texts are culturally neutral and at the B1 (intermediate) level, as defined by the Common European Framework of Reference for Languages (Council of Europe, 2001). We made slight adjustments to ensure uniformity, resulting in four texts of approximately 200 words each, containing 16 or 17 sentences per text (66 sentences in total). The English texts and their Mandarin translations are provided in Appendix B. The translation process followed the same protocol as for the word lists, with native speakers responsible for both the initial translation and subsequent review. Translators were instructed to avoid overly literal or unnatural phrasing while staying as close as possible to the original English text in terms of vocabulary and word order, within the syntactic constraints of the target language. This approach aimed to produce texts that were as comparable as possible across the different languages. Finally, each word in the corpus was tagged with Part of Speech (PoS or lexical category) information (see Section 2.2). The set of PoS tags that we used is provided in Table 1. However, tokens annotated with 'intj' were not processed.

Table 1. PoS tags assigned to the word tokens in the European and Chinese texts.

PoS Tag	Description
$	sentence boundary
noun	noun
verb	verb
adj	adjective
det	articles, demonstrative and possessive pronouns, quantifiers
pron	pronoun, all other pronouns
num	cardinal number
adv	adverb
adp	adposition (preposition, postposition)
con	conjunction (coordinating or subordinating)
intj	interjection
part	particle

2.2. Measuring Linguistic Distances

Given 17 European and 15 Chinese varieties, we have a total of 32 varieties. We compared each variety to each variety at both the lexico-phonetic and the syntactical level. In the subsections below, we explain how the distances were measured.

2.2.1. Lexico-Phonetic Distances

As explained in Section 2.1.2, we obtained, for each variety, phonetic transcriptions of the translations of a set of 100 English words. Therefore, when comparing the lexico-phonetic distance between two varieties, 100 word pairs are considered. For each pair, the Levenshtein distance is calculated between the two word transcriptions, regardless of whether the words are cognates of each other. Kessler (1995) refers to this as "all-word distances", in contrast to "same-word distances", which are measured on cognate word sets only. The aggregated distance between two varieties is the average of 100 word pair distances.

The Levenshtein distance is defined as the penalty incurred by the least costly set of edit operations (insertion, deletion and replacement of a symbol) needed to convert the string of phonetic symbols in language A to its counterpart in language B. The basic idea of the Levenshtein distance is illustrated in the example shown in Table 2. English *children* can be changed to the German cognate *Kinder* by replacing [ʧ] by [k] and [l] by [n], deleting [ɪ], [ə] and [n], and by inserting [ɜ]. As can be seen in the table, the normalized Levenshtein distance between the two realizations is 67%. The normalization divides the raw cost by

the maximum cost that could be incurred given the strings under comparison. This is done since long strings provide more opportunities for discrepant symbols than short strings.

Table 2. Alignment of the realization of English *children* and German *Kinder*. The normalized Levenshtein distance is 4.0/6.0 = 0.67 or 67%.

Steps	**Alignment Slots**								**Total**
English	ʧ	ɪ	l	d		ɹ	ə	n	
German	k	ɪ	n	d	ɜ				
Actual cost	1		1		0.5	0.5	0.5	0.5	4.0
Maximum cost	1	1	1	1	0.5	0.5	0.5	0.5	6.0
Normalized cost									4.0/6.0 = 0.67

The transcriptions of the two realizations can be aligned to each other in many different ways, but the Levenshtein distance always gives the cost of the cheapest mapping. We used VC-sensitive alignment, i.e., a version with the constraint that vowels may match only with vowels, and consonants only with consonants. Exceptions are /j, i, w, u/ which may align with any segment, and [ə] and [ɐ], which may be aligned with any vowel or sonorant consonant.

We used Pointwise Mutual Information (PMI) Levenshtein as a method for comparing dialects (Wieling et al., 2009; Wieling, 2012).[4] This version of the Levenshtein distance learns the segment distances by analyzing the alignments (such as in Table 1) that underly the distance measurements. The basic idea is that substitutions of segments that frequently co-occur in an alignment slot are weighed less heavily than segments that rarely co-occur. These segment distances are used as operation weights, rather than the categorical weights that we used in the example in Table 2. As a result, [i] and [ɒ], as an example of an unlikely substitution, will be more distant to each other than [i] and [ɪ]. We assume here that data-driven weighting based on the actual occurrence of sound correspondences between a pair of languages should yield a psychologically more realistic estimate of the effort needed by a native listener to crack the code of the other language, more so than a feature weighting based on general phonetic principles (as, e.g., in Heeringa, 2004).

Following Heeringa et al. (2023), differences in length between segments were processed by adapting the phonetic transcriptions before calculating the Levenshtein distances as follows:

- If a segment is transcribed as extra short (e.g., [ă]), it remains unchanged;
- If a segment does not have any length mark, it is doubled, e.g., [a] becomes [aa];
- If a segment is marked as half-long, it is tripled, e.g., [aˑ] becomes [aaa];
- If a segment is marked as long, it is quadrupled, e.g., [aː] becomes [aaaa].

As in Heeringa et al. (2023), a small set of diacritics was processed. These are listed in Table 3. An aspirated sound such as [tʰ] is processed by treating it as half a [t] and half an [h]. When comparing [tʰ] with, e.g., [s], the distances between [t] and [s] and between [h] and [s] are calculated, and then averaged. The other diacritics are processed in a similar way.[5]

We decided not to process properties of word prosody even though words may differ in the location of the stress (whether primary or secondary) and in the word tone. Swedish (Gårding, 1974; Bruce, 1977; Riad, 2014) and Croatian (Inkelas & Zec, 1988) have a binary word-tone opposition (only in stressed syllables) while the Chinese varieties have four (Mandarin varieties), five or six (non-Mandarin varieties) lexical tones (e.g., Yip, 2002; Mian Yan, 2006). The reason we did not process word prosodic features is that we do not know how to weigh word prosodic differences against segmental differences. Moreover, although several attempts have been made to quantify tonal differences in a mean-

ingful way (e.g., Mongeau & Sankoff, 1990; Yang & Castro, 2008; Tang & van Heuven, 2015), lexical tone information made no contribution to the prediction of either cross-dialectal intelligibility or of cladistic distance in the genealogical trees of Sinitic languages (Tang & van Heuven, 2015).

Table 3. Diacritics processed when calculating phonetic distances.

Diacritic	Example	Averaged with
aspirated	tʰ	[h]
labialized	tʷ	[w]
palatalized	tʲ	[j]
velarized	tˠ	[ɣ]
pharyngealized	tˤ	[ʕ]
nasalized	ã	[n]

2.2.2. Syntactic Distances

For measuring syntactic distances, we use the PoS tag n-gram method that was introduced by Nerbonne and Wiersma (2006), who used this method in order to measure the total impact of L1 on L2 syntax in second language acquisition on the basis of a corpus of English collected from Finnish Australians. While Nerbonne and Wiersma used the method for comparing accents of English, the method can be used for comparing any pair of language varieties, when for each variety a text is available in which each word has been assigned a Part of Speech (PoS) tag.

The procedure is as follows: First, an inventory of n-grams of PoS tags across the texts of the different language varieties is made. Then, the number of occurrences for each n-gram per variety is counted. Thus, we obtain a vector of n-gram counts for each variety. The syntactic distance between any two varieties is then calculated by comparing their respective frequency vectors. According to Nerbonne and Wiersma (2006, p. 85), the "choice of vector difference measure (…) does not affect the proposed technique greatly, and alternative measures can be used straightforwardly". Di Buccio et al. (2014) used the cosine similarity, i.e., angle θ between the vectors. Swarte (2016) and Heeringa et al. (2018) used Pearson's product-moment correlation coefficient r. Both similarity measures are easily converted to distance measures by calculating, respectively, $1 - \theta$ and $1 - r$. An advantage of either measure is that they are insensitive to differences in scale. This is important because the raw n-gram frequencies will increase with the number of words in the text. As a result, frequency vectors may have different scales, but with the above measures, this difference in scale is normalized. Another advantage is that both measures range between 0 (no difference) and 1 (maximum difference) and are, therefore, easy to interpret.

We illustrate this method using an example. Assume the sentence and its annotation as given in Table 4. From this sentence, nine PoS tag trigrams can be extracted, as shown.

Once all sentences of a text are analyzed in this way, the frequencies of the n-grams (here: trigrams) in the text can be counted. If we do this for texts of multiple varieties, we can calculate distances between the texts by comparing the respective frequency vectors with each other by means of the cosine measure or Pearson's correlation coefficient.[6]

Table 4. Trigrams extracted from an example sentence annotated with PoS tags.

	English	is	the	most	spoken	language	in	the	world	
	noun	verb	det	adv	adj	noun	adp	det	noun	
$	noun	verb								
	noun	verb	det							
		verb	det	adv						
			det	adv	adj					
				adv	adj	noun				
					adj	noun	adp			
						noun	adp	det		
							adp	det	noun	
								det	noun	$

2.3. Visualization

2.3.1. Cluster Analysis

We apply hierarchical cluster analysis to the linguistic distances among the 32 varieties. This results in a hierarchically structured tree, which is called a dendrogram. In a dendrogram, the leaves represent the varieties and the branches represent the distances between the varieties and the clusters (Jain & Dubes, 1988). We used the Unweighted Pair Group Method using Arithmetic averages (UPGMA).

The UPGMA algorithm generates rooted dendrograms under the assumption of a constant rate of change, implying an ultrametric tree where the distance from the root to each branch tip is the same. When studying historical relatedness or divergence, this assumption may be problematic. After all, languages may change unevenly (due to contact, socio-political pressure, isolation, etc.). Some languages may undergo rapid shifts, while others remain relatively stable. Therefore, scholars who aim to reconstruct the historical development of the languages prefer to visualize linguistic relationships using a Neighbor-Joining (NJ) tree, which does not assume a constant rate of change.

Our primary objective, however, is rather to interpret linguistic distances as a proxy for mutual intelligibility. As illustrated in Section 1, a Cantonese speaker and a Mandarin speaker who each understand only their own variety would not be able to communicate with one another. In this sense, the two varieties are perceived as distinct languages—not as a result of historical divergence, but rather due to their lack of mutual intelligibility arising from numerous structural differences.

Given this perspective, we opt for UPGMA, as the distances represented in the dendrograms typically account for the highest proportion of variance in the original distance matrix—see Sokal and Rohlf (1962). Specifically, for lexico-phonetic distances, the UPGMA tree accounts for 95.9% of the variance in the original data, while for syntactic distances, it explains 94.5% of the variance.

2.3.2. Multi-Dimensional Scaling

In order to make sure that the results that are obtained with UPGMA clustering are robust, we visualize the mutual distances among the 32 varieties by means of multi-dimensional scaling as well. Using this technique, the 32 varieties are plotted in a two-dimensional diagram so that the distances suggested by the diagram reflect the original distances as closely as possible (Torgerson, 1958). There exist several alternatives. We use classical multi-dimensional scaling, because in our case, this procedure produces a graph in which the distances between the varieties best match the original distances. Addition-

ally, the main groups found in the UPGMA trees are projected on the multi-dimensional scaling plots.

3. Results

On the basis of the measures explained in the previous section, we computed four distance matrices. These can be consulted in Appendix C for the 17 × 17 European language pairs and in Appendix D for the 15 × 15 Chinese varieties. The distances we computed are symmetrical, i.e., the distance from variety A to variety B is equal to the distance from B to A. The upper triangle of the matrix displays the lexico-phonetic PMI-weighted Levenshtein distances, while the lower triangle shows the syntactic distance computed on the PoS trigram frequencies.

In the following subsections, we will not refer to the distance tables but limit the presentation to the tree diagrams and MDS plots that can be derived from the distance matrices. Note that the diagrams and plots were computed on the combined matrices for the European and Chinese varieties, i.e., the 32 × 32 square matrix, even though the distances between pairs of varieties in Appendices C and D are specified separately for European and Chinese varieties. As will be shown in the following subsections, the distance between languages varieties across the European–China divide are always larger than between language pairs within the sets, so that the primary split in the tree diagrams will be between European versus Chinese varieties—both in terms of the lexico-phonetic distance and syntactic distance. We will first present the lexico-phonetic distances (Section 3.1), and then deal with the syntactic distances (Section 3.2).

3.1. Lexico-Phonetic Measurements

Figure 1 shows the affinity tree based on the PMI-weighted Levenshtein and normalized distances between all pairs of the 32 × 32 language varieties included in the study. The UPGMA clustering accounts for 95.5% of the variance.

Predictably, the first split in the tree is in a European versus Sinitic cluster. The internal structure of the European languages reflects the tripartite division into the Germanic, Romance and Slavic groups, where the Germanic and Romance groups are closer to one another than either is to the Slavic group. Within the Germanic group, Dutch and Frisian are most similar, then joined by German. On the next-higher level, this cluster is joined by the Scandinavian (Danish–Swedish) pair, while English is least similar to all the continental Germanic languages. These results have been published before (e.g., Heeringa et al., 2023) but the new element here is the inclusion of Frisian. Lexico-phonetically, Frisian is most similar to Dutch, rather than to English or Danish.[7]

The most similar pair of Romance languages is Spanish–Italian. No other clusters are seen. Other Romance languages are added to the clustering one by one, first Portuguese—suggesting a cluster of South Romance languages—then Romanian, and last, French.

Czech and Slovak are the two closest Slavic languages in our sample. This is expected, since until recently, there was only one common language serving the Czechoslovak nation, before it split up into two separate countries in 1993. The pair forms a West-Slavic subcluster with Polish. The remaining three languages form the South Slavic subcluster, in which Slovene and Croatian are more similar to each other than to Bulgarian.

The Chinese branch of the lexico-phonetic tree shows that Wuhan and Chengdu are the most similar pair of Mandarin dialects (South-Western Mandarin), followed by the Xi'an–Jinan Mandarin pair, which is joined by Taiyuan at the next stage in the tree, forming the Northern Mandarin group. Unexpectedly, the Beijing dialect is joined to the South-Western group rather than to the Northern Mandarin subgroup. Nevertheless, all Mandarin varieties are grouped into one overarching cluster. In the non-Mandarin (Southern)

Sinitic dialects, the pair of Wu dialects (Suzhou–Wenzhou) is the most similar. Nanchang and Meixian form a subcluster, which is joined by Guangzhou at the next level. Unexpectedly, there is no cohesive cluster of Min dialects. Finally, Changsha dialect, although traditionally classified as a Southern Sinitic dialect, ends up in the Mandarin branch of the Sinitic subtree.

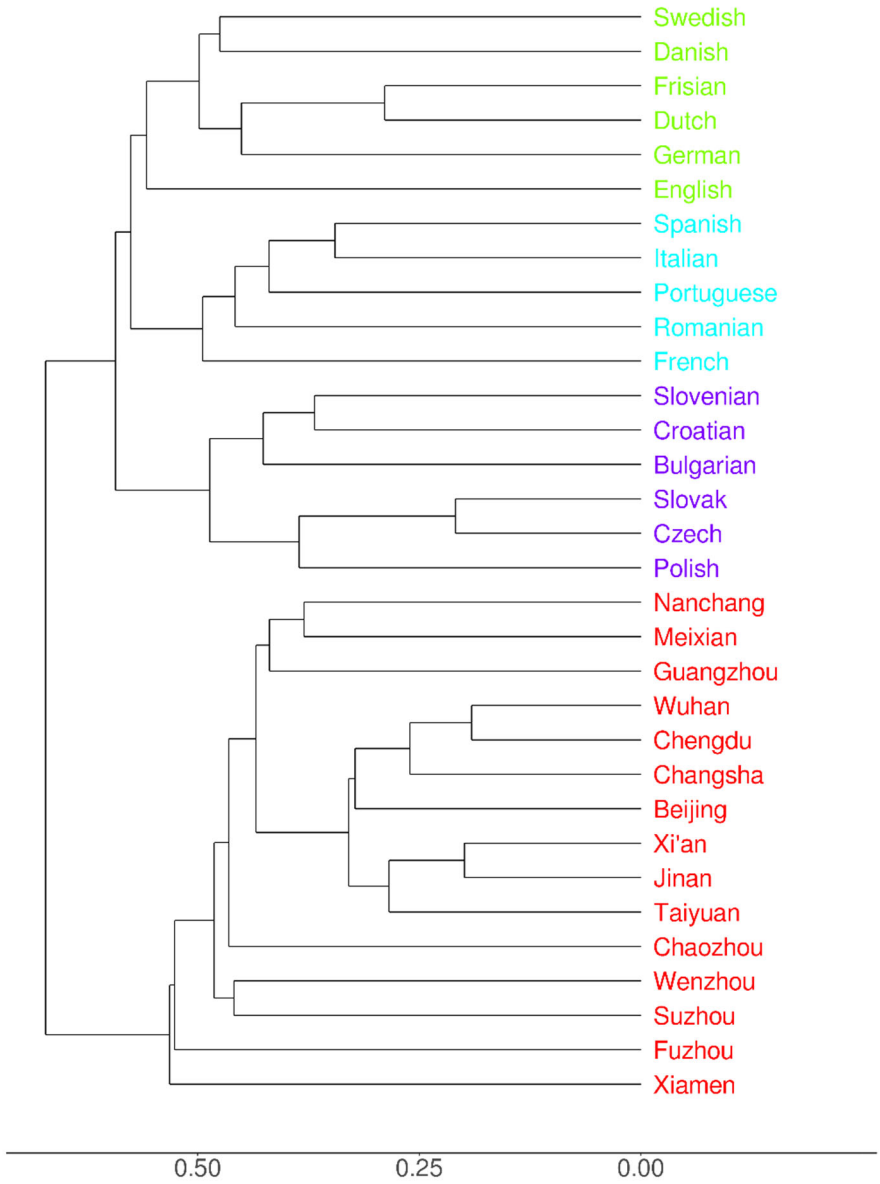


Figure 1. Dendrogram for all 32 language varieties, derived from PMI-weighted lexico-phonetic Levenshtein distances (using UPGMA clustering). Language families are color coded: green for Germanic, blue for Slavic, purple for Romance and red for Sinitic.

Figure 2 presents the (classical) MDS plot based on the same lexico-phonetic data as the tree diagrams in Figure 1.

The two dimensions in the MDS account for 73.3% of the variance in the distances (Kruskal's stress-1: 0.499, i.e., a bad fit). The dimensions are difficult to interpret. Dimension 1 separates the three European language families (positive values) from the 15 Sinitic dialects (negative values), leaving a wide gap between the two. Most likely, this is a reflection of the complete absence of lexical overlap between the two groups of languages. Dimension 2 separates the Slavic languages (positive values) from the Romance and Ger-

manic languages (negative values), which still form cohesive clusters but with relatively little separation—which reflects the tree structure in Figure 1.

The Chinese dialects have neutral values (close to 0) along Dimension 2, and therefore find themselves in between the Slavic cluster (to the right) and the Romance–Germanic groups to the left. However, the Chinese dialects do not vary much along Dimension 2. Most of the variation among the Chinese dialects is seen in Dimension 1, with the six Mandarin dialects (plus Changsha, as in the dendrogram) at the bottom of the plot, while the non-Mandarin varieties are characterized by less extreme values along this dimension.

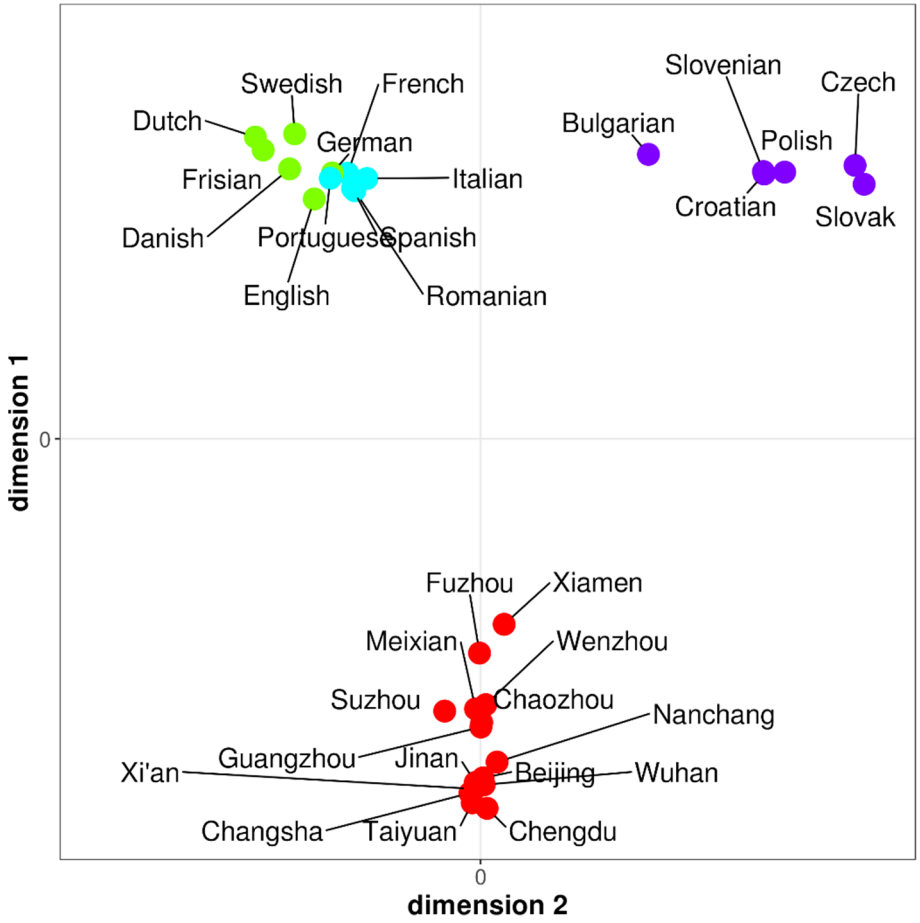


Figure 2. Classical MDS plot based on PMI-weighted Levenshtein lexico-phonetic distances among 17 European languages (green for Germanic, blue for Romance, purple for Slavic) and 15 Chinese dialects (red dots).

Figure 2 suggests that the lexico-phonetic distances in the Romance family are small in comparison to the Germanic languages, while the Slavic languages show the largest dispersion along Dimension 2 (while the Dimension 1 value remains approximately the same across all European languages). A similar observation can be made for the Chinese dialects, if we exchange the dimensions. Here, the Mandarin dialects (plus Changsha) form a tightly clustered group, while the non-Mandarin dialects are dispersed over a wider range of Dimension 1 values. The total dispersion along D1 for the Chinese dialects is of a similar magnitude to the dispersion of the Slavic languages along D2. However, the dispersion of the Germanic, and especially Romance languages, is very small in comparison.

3.2. Syntactic Measurements

Figure 3 contains the dendrogram computed for the total sample of 32 language varieties based on the correlation of the PoS trigram frequencies, which arguably capture

the similarity in the syntactic surface structure (word order) among the varieties. The UPGMA clustering captures 94.5% of the total variance in the syntactic distances among the 32 varieties.

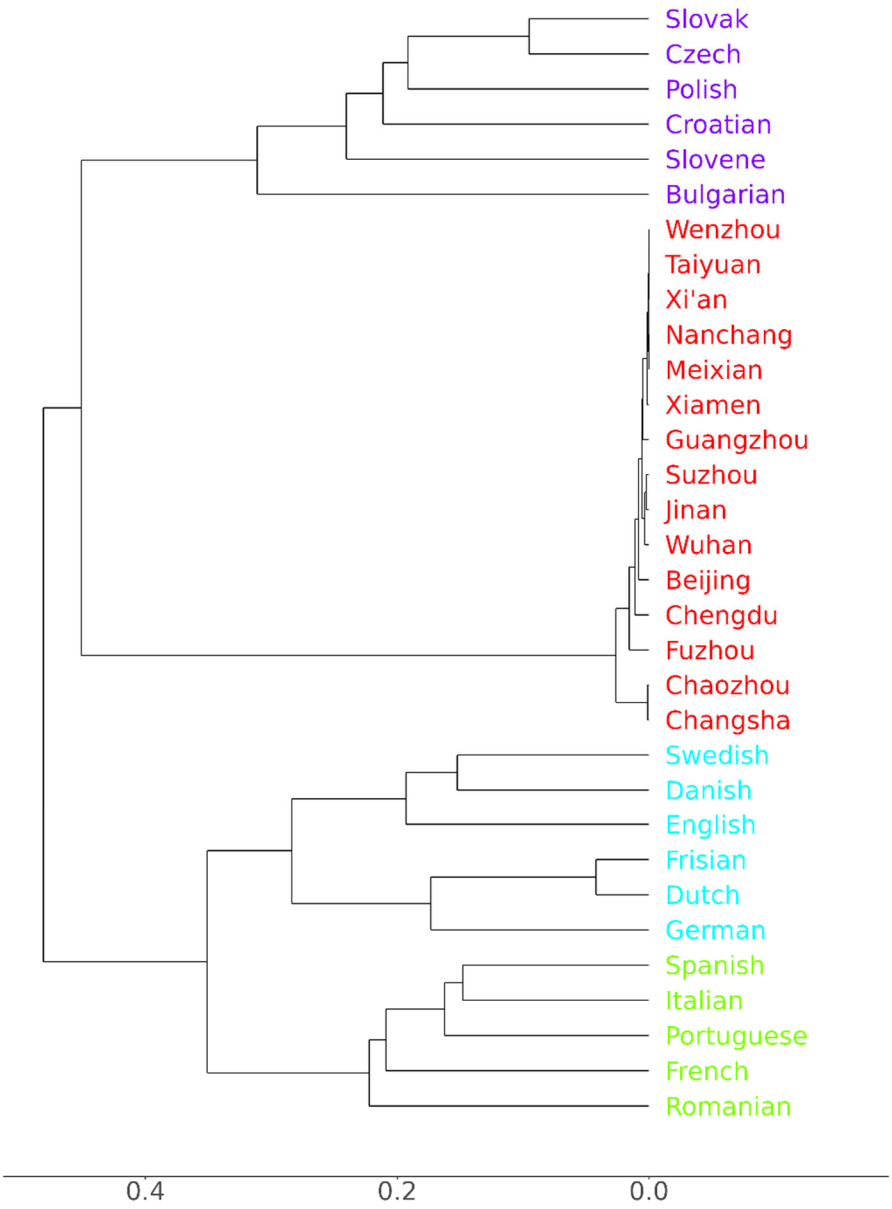


Figure 3. Dendrogram for all 32 language varieties, derived from syntactic distances based on correlation of PoS trigram frequencies (using UPGMA clustering). Language families are color coded: green for Germanic, blue for Slavic, purple for Romance and red for Sinitic.

On surface-syntactic grounds, the 32 varieties are readily grouped into four cohesive clusters, each corresponding to one of the four families involved in the present study. This mirrors the results seen for the lexico-phonetic distances in Figure 1. In terms of lexico-phonetic distance, however, the European languages form one primary cluster against the Chinese dialects, while in the syntactic dendrogram, Romance and Germanic languages are one primary cluster while the Slavic cluster is parsed together with the Chinese cluster.

It is obvious in Figure 3 that the 15 Chinese dialects are syntactically highly similar, in some cases even identical, in contradistinction to the European languages, all of which differ more from each other than any difference among the Chinese varieties. The syntactically closest European language pair is Dutch–Frisian with a distance of 0.042, while the

second-closest pair is Czech–Slovak (distance: 0.095). The syntactically most distant pair of Chinese dialects is Changsha–Fuzhou, with a distance of 0.038—which is still less than the distance between Dutch and Frisian.

The internal structure of the dendrograms for each of the three European language families is roughly the same, whether based on the lexico-phonetic or on the syntactic trigram distance. There are only small discrepancies between the two domains for Germanic and Slavic, while the very same tree structure is seen in the Romance family. The strong correlation between the domains was observed earlier by Heeringa et al. (2023). In the earlier study, no distances were computed between language pairs across family boundaries. The present study includes cross-family comparisons, and shows that the languages within each European family are so similar to one another and so different from any language in the other families that perfectly separated clusters are found.

Figure 4 shows the distances among the 32 languages in our sample in an MDS plot. The reduction of the distances to two dimensions still explains 91.0% of the original variance (Kruskal's stress-1: 0.196, i.e., a fair fit). The plot shows roughly the same topography as in Figure 2 for the lexico-phonetic distance.

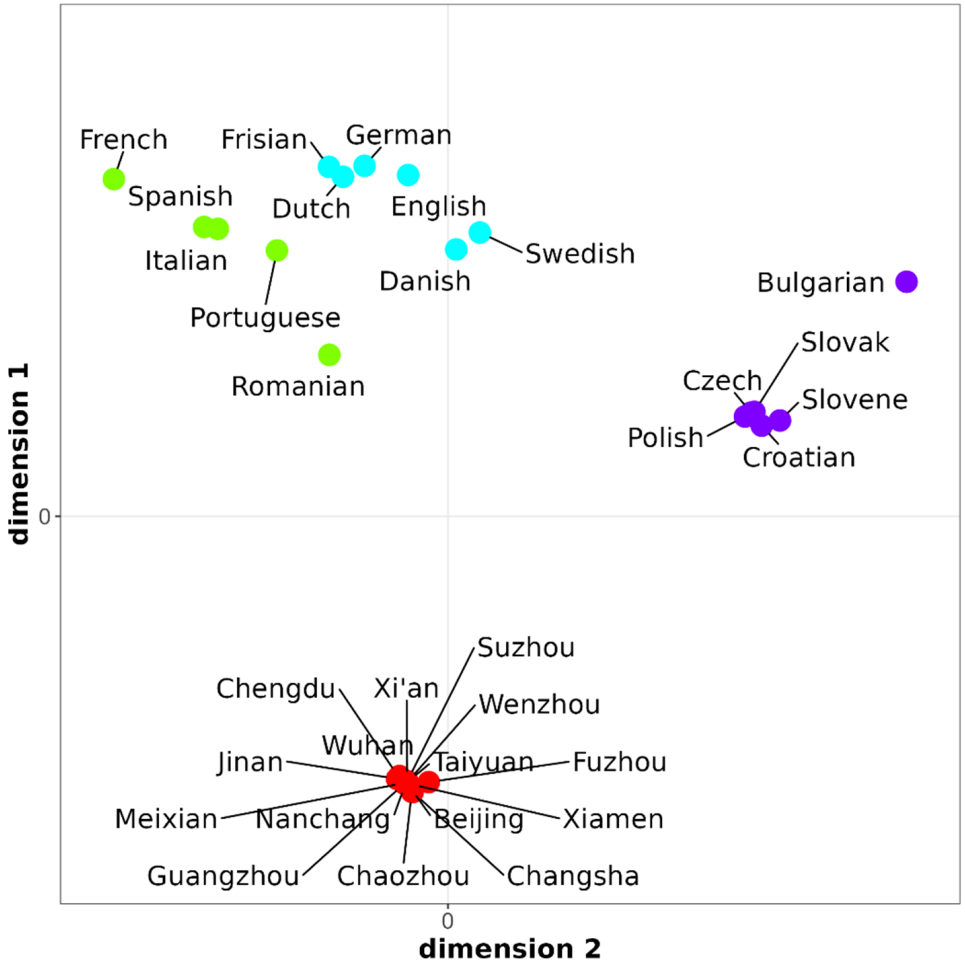


Figure 4. Classical MDS plot based on syntactic distances based on correlation of PoS trigram frequencies among 17 European languages (green for Germanic, blue for Romance, purple for Slavic) and 15 Chinese dialects (red dots).

Again, the Chinese dialects have strong negative values on Dimension 1 against high positive values for the European languages. Dimension 2 separates the three European language families into Romance (left), Germanic (mid) and Slavic (right). The plot also shows that the 15 Chinese dialects are syntactically very similar, as is foreshadowed in

Figure 3. The distance between the Chinese dialects as a group and each of the three European language families is larger than the distance between the European language families. Importantly, the Slavic family as a whole is closer to the Germanic language group than to the Chinese dialects.

Within the Germanic group, Dutch and Frisian are the nearest one another, followed by Danish and Swedish. German is closest to Dutch and Frisian, while English is somewhat closer to these three than to the Scandinavian pair. In the Romance group, Spanish and Italian are closest together, followed by Portuguese. Romanian and French assume extreme positions on opposite sides within the Romance group. In the Slavic group, Bulgarian is rather remote from the other five languages.

The Chinese dialects are so close together in Figures 3 and 4 that we present two additional figures, which show only the 15 dialects with increased resolution, thereby zooming in on the small differences in syntactic distance among the dialects. Figure 5 presents the high-resolution dendrogram for the Chinese dialects only. The UPGMA clustering yields a tree structure that captures 92.0% of the original variance in syntactic distances.

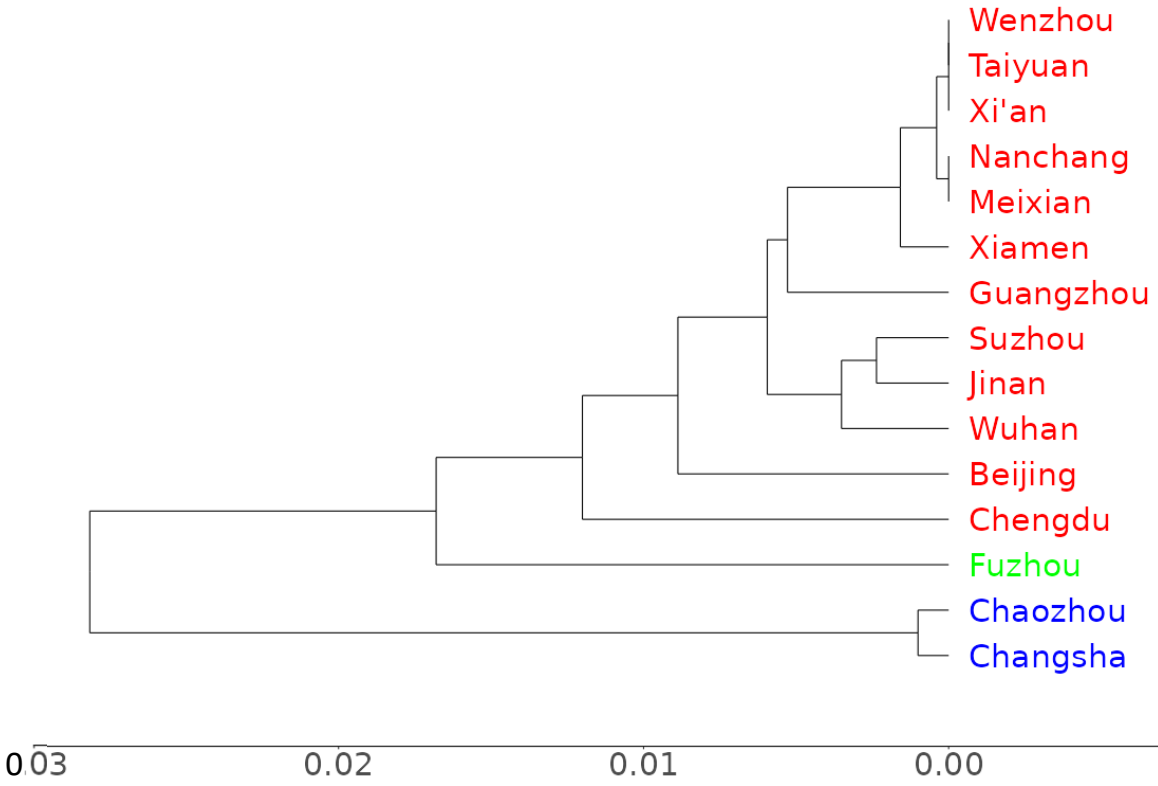


Figure 5. Dendrogram with expanded scale for only 15 Chinese dialects, derived from syntactic distances based on correlation of PoS trigram frequencies (using UPGMA clustering). Non-red labels mark most deviant varieties.

The tree structure bears no resemblance to the traditional genealogy of the Chinese dialects. Taiyuan and Xi'an, two Northern Mandarin dialects, are one cluster with zero syntactic distance, but do not differ from Wenzhounese, which is a dialect in the Wu group and should be in a cluster with the other Wu dialect, i.e., Suzhou—but is not. Similarly, Mandarin Jinan and Wu Suzhou form a cluster at the lowest level—again across the Mandarin–Southern divide.

An enlarged MDS plot, as in Figure 6, nevertheless shows that the small syntactic distances among the 15 Chinese dialects do not vary completely at random but do reflect genealogical relationships, at least to some extent. The classical MDS scaling, reducing

the syntactic trigram distances to two dimensions, here accounts for 93.4% of the original variance (Kruskal's stress-1: 0.168, i.e., a fair fit).

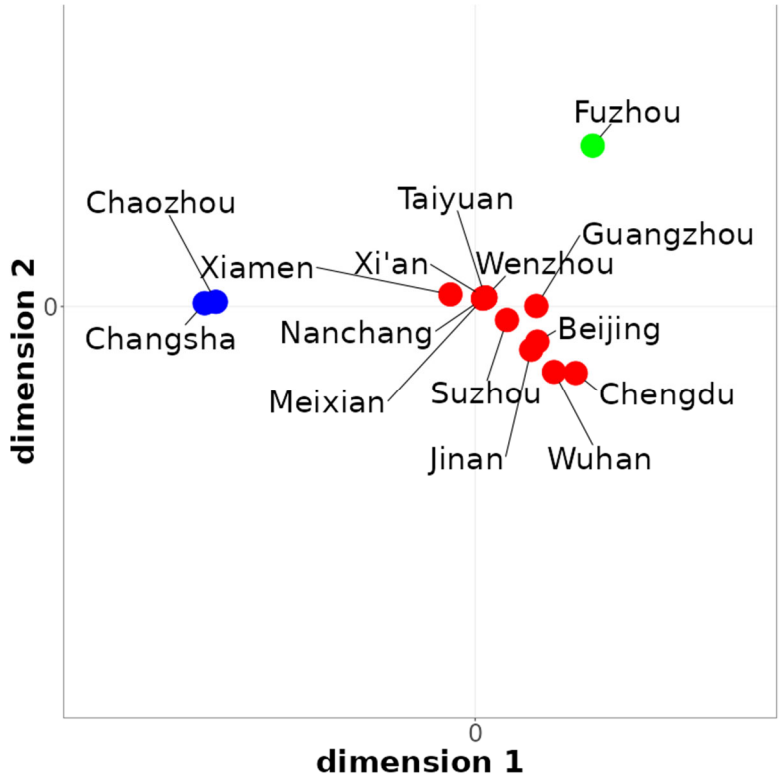


Figure 6. Classical MDS plot with expanded scales for only 15 Chinese dialects, derived from syntactic distances based on correlation of PoS trigram frequencies. Non-red dots mark outlier varieties.

In the MDS plot, the six Mandarin dialects are close together, with the two Southern Mandarin dialects (Wuhan, Chengdu) at the bottom-right and the Northern Mandarin dialects more towards the origin (0, 0) of the space. Unexpectedly, however, in the middle of the Northern Mandarin quartet, there is the Wu pair Suzhou–Wenzhou. There are several more 'intruders' in the Mandarin group, suggesting, for instance, that Mandarin Beijing and Cantonese (Guangzhou) are syntactically very similar. Fuzhou, Chaozhou and Changsha seem relatively remote from all other dialects, indicating that the periphery of the plot is occupied by non-Mandarin dialects only.

Inspection of the configuration of data points, especially in Figures 2 and 4, reveals substantial differences among language groups in the spread of the individual varieties around the family centroid in the MDS plots. For instance, in terms of lexico-phonetic distance, the varieties within the Germanic and in the Romance families are closer to their centroids than the Slavic varieties. Similarly, it was observed above that the Chinese varieties are closer together in terms of their syntactic structure than the languages in the European families. To show the differences in the tightness of the clusters in the MDS plots, we compute the Euclidean distance for each language variety from the respective family centroid, and plot these in Figure 7. This is performed for lexico-phonetic and for syntactic distance separately, while the varieties are split into five rather than four families by grouping the Mandarin and Southern Sinitic varieties into distinct subfamilies. Figure 7 shows that the lexico-syntactic distances within the European language families are roughly equal to the Sinitic language varieties. Within the European families, the smallest distances are seen in Romance, which are three times smaller than in the other two families. In the Sinitic group, the Mandarin languages have three-times-smaller distances among them than the

Southern languages. Importantly, the lexico-phonetic distances within the Romance and Mandarin groups are about the same.

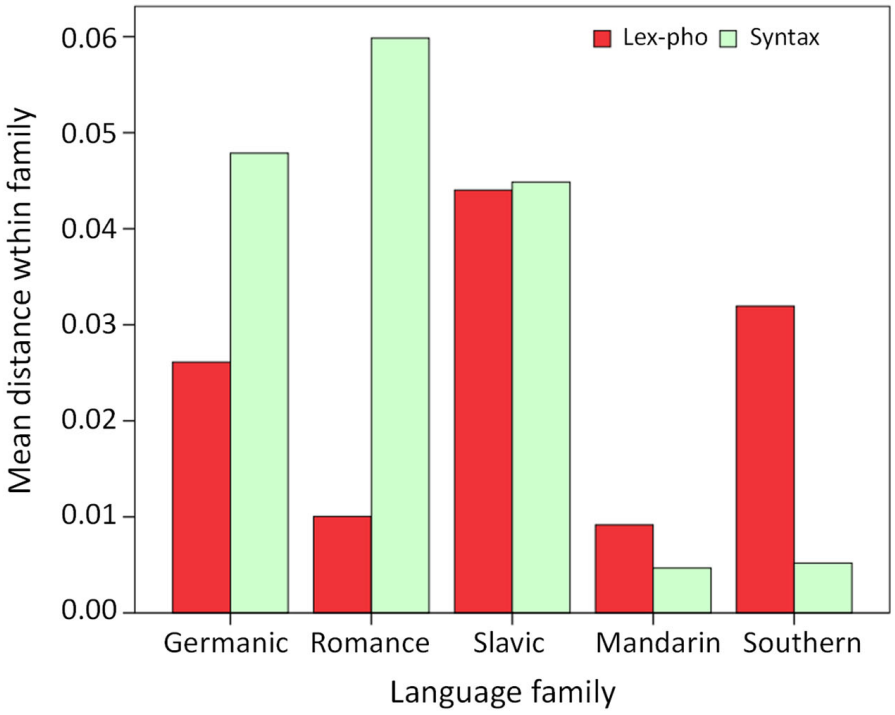


Figure 7. Mean (Euclidean) distance among language varieties in each of five language (sub)families.

The syntactic distances within the European families are relatively large, roughly ten times larger than the syntactic distances within the Sinitic language groups—which are equally small in the Mandarin and Southern groups.

Table 5, finally, lists the lexico-phonetic and syntactic distances, both raw as in Appendices C and D and computed as in the MDS plots above for the four specific pairs of language varieties that are explicitly mentioned in the quotations in the introduction.

Table 5. Lexico-phonetic and syntactic distance between members of four selected language pairs.

Language Pair	Raw Distance		MDS Distance	
	Lexico-Phonetic	Syntactic	Lexico-Phonetic	Syntactic
Mandarin–Cantonese	0.473	0.009	0.041	0.005
Portuguese–Romanian	0.479	0.199	0.022	0.091
Portuguese–Italian	0.427	0.162	0.029	0.049
German–Dutch	0.449	0.174	0.068	0.019

3.3. Correlating the Domains

It was observed in the preceding sections that the lexico-phonetic and syntactic distances are correlated in the European languages (compatible with earlier findings by Heeringa et al., 2023) but not in the Sinitic languages. To substantiate these rather informal observations, we correlated the lexico-phonetic distances with the syntactic distances, for both the Chinese (120 unique and non-identical language pairs) and the European (136 pairs) language groups, as seen in Figure 8. Indeed, for the Chinese language group, no significant correlation was found. For the European language group, a significant correlation was found of $r = 0.690$.[8]

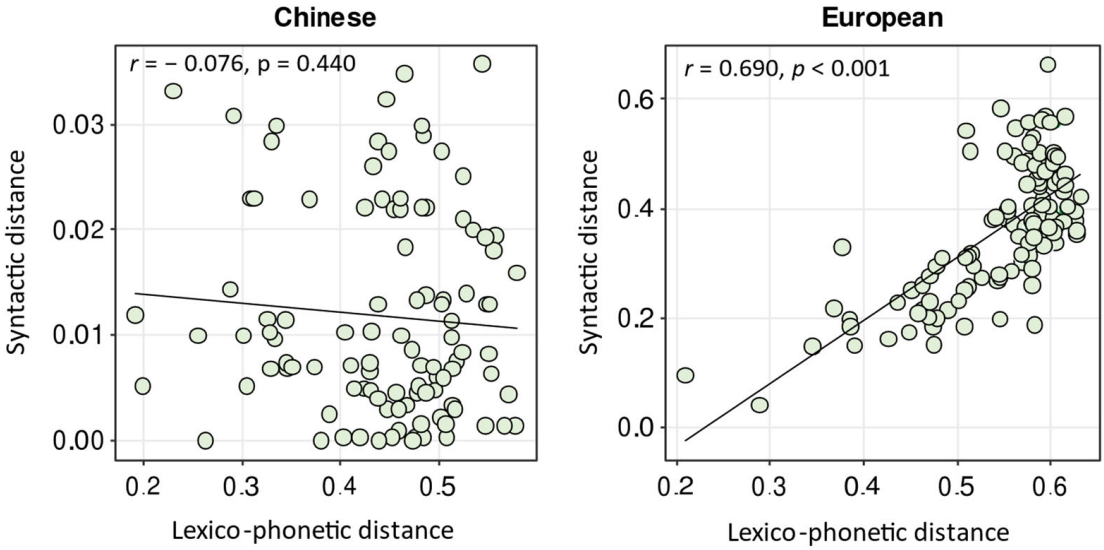


Figure 8. Correlation between the lexico-phonetic distances and syntactic distances for Chinese (**left panel**) and European (**right panel**) language varieties.

4. Comparison and Discussion

4.1. Comparison with Published Affinity Trees

The lexico-phonetic similarity among European languages was investigated earlier by Polyakov et al. (2009) using the ASJP data on 40 Swadesh concepts (see Section 1). Their sample of European languages overlaps only partially with ours, i.e., three shared Germanic languages, three Romance and three Slavic languages. As in our own results, the three groups end up in three different branches of the affinity tree. Moreover, Danish and Swedish cluster at the lowest level in Germanic, then are joined by Dutch. Portuguese and Italian cluster first, to be joined next by French in the Romance branch, while Czech and Polish cluster before being joined by Bulgarian in the Slavic group. This is also seen in our own trees. However, "borderline" languages such as English (Germanic but with heavy influence from Romance) and Romanian (Romance but strong Slavic influence on vocabulary) are not included by Polyakov et al. Computations done by ourselves using our PMI-weighted Levenshtein Distance measure and UPGMA clustering on the ASJP data on the web show a correlation with our own lexico-phonetic distances at $r = 0.925$.

We also (and perhaps more appropriately) compare our results with the fuller tree computed by Müller et al. (2021), which contains all the European languages covered in the present paper (and many more), as well as 13 of our 15 Sinitic varieties. The methodology was basically the same as in Polyakov et al. (2009), again based on a comparison of the 40 Swadesh items.

Müller et al. (2021) add German as the most distant language to a cluster of two pairs (Danish + Swedish and Dutch + Frisian), whereas our tree agrees with the more widely accepted view that English differs most from the continental Germanic varieties. The tree for the five Romance languages has the same add-on structure as in our results. For the six Slavic languages, we would argue that our results reflect the common division of varieties more closely than those reported by Müller et al. We have a primary split in West Slavic and South Slavic, whereas Müller et al. have Bulgarian as a late add-on to a hybrid cluster of the other five varieties (see Appendix E for a side-by-side comparison of the affinity trees we extracted from Müller et al. and our own).

As for the Sinitic varieties, one Mandarin and one non-Mandarin variety in our set are not covered by Müller et al. (2021), i.e., Mandarin Taiyuan and non-Mandarin Chaozhou (Wu). A comparison of the lexico-phonetic affinity tree that can be extracted for the

Sinitic varieties from Müller et al. can be found and graphically compared with ours in Appendix F. The structure of the Müller tree is not unlike ours but there are important differences. Our tree keeps the two Wu varieties (Suzhou, Wenzhou) together, while these are scattered in the Müller tree: Wenzhou is added to the predominantly Mandarin cluster, while Suzhou is clustered with Meixian (Hakka). Also, our tree shows a less polluted Mandarin cluster with one non-Mandarin intruder (Changsha) against two intruders in the Müller tree (Changsha and Nanchang). The Müller tree is better in terms of its internal structure of the Mandarin cluster: it keeps Beijing together with the other Northern varieties, while it is parsed with the SW group in our tree (although this may be due to the elimination of SW Taiyuan from our tree). On aggregate, our lexico-phonetic distances seem to reflect the traditional phylogeny (slightly) better than the Müller et al. solution.

4.2. PoS Trigram Frequencies and Word Order Typology

We noted, in Section 3.2, that the syntactic distances among the Sinitic varieties were found to be roughly ten times smaller than those among the European varieties. The syntactic distance was expressed as the complement of Pearson's correlation coefficient computed on the PoS trigram frequencies counted on our mini-corpus of parallel translations of four relatively simple texts, with a total length of about 800 words (for the English originals, see Appendix B). The Part of Speech (PoS) categories we used were relatively broad but—as a consequence—applicable to all 32 language varieties in our sample. We have no reason to assume that the ten labels we employed could be biased in favor of the European languages. They should apply equally well to the Sinitic varieties. The reason why the European languages vary so much more than the Sinitic varieties should therefore be sought in the existence of substantial and pervasive word order differences in our 17 European languages, and the absence of such differences among the 15 Sinitic varieties we included. The varieties were not selected with particular syntactic properties in mind. They were a legacy from the first author's doctoral dissertation (Tang, 2009) on the experimentally measured cross-lingual intelligibility of Mandarin and non-Mandarin Chinese varieties, and the prediction thereof from phonetic and lexical similarity.

That the European languages should differ substantially in terms of our trigram frequencies can be expected from even a few well-known basic differences in word order. One is the habit of Germanic languages to string nouns together to compounds (*steam train*), where Romance languages break such compounds down into a series of preposition phrases (*train* à *vapeur*). Consequently, we would expect noun–prep–noun trigrams to be more frequent in Romance than in Germanic languages. Similarly, Romance languages have the qualifying adjective following the noun (*vin blanc*) rather than preceding it as in Germanic languages (*white wine*). As a consequence, det–noun–adj trigrams should be frequent in Romance while det–adj–noun trigrams should abound in Germanic languages. Third, most Germanic languages allow constituents between auxiliaries/modals and the corresponding participle or infinitive. In Romance languages, the verb group cannot be separated. Accordingly, X–verb–verb and verb–verb–X trigrams will be more frequent in Romance than in Germanic languages (with the exception of English, which keeps its verbs together). We consider it beyond the scope of the present article to trace the differences in trigram frequencies back to specific typological features, nor will we engage in actually testing these predictions here. Validating the trigram method will be a topic for a future paper. We reiterate that the trigram method works well for the European languages we included; the syntactic distances correlate well with the lexical and phonetic distances computed for the same languages (Heeringa et al., 2023). We assume they also work well for the Chinese varieties.

The reason, then, that the syntactic distances among our 15 Chinese varieties are small must be that these varieties tend to accept the same word orders. We will not attempt to run an analysis of typological word order differences and similarities in our Sinitic varieties, even though the World Atlas of Linguistic Structures (WALS, Dryer & Haspelmath, 2013) lists a plethora of word order features—this, again, will await a future study. For now, we rely on observations that can be found in the literature on word order differences across Sinitic varieties.

A comprehensive literature review is given by Z. Zhang (2003). His conclusion is that all Sinitic varieties have the same basic word order, i.e., SVO (subject–verb–object). The SVO order is broken down into two sections: the SV part and the VO part. In a number of Sinitic varieties, the first part has the reverse VS. Similarly, the second part is seen as OV in a number of varieties. Importantly, however, none of the deviant varieties overlap with our sample of 15. Moreover, deviations from SVO are typically found in marked sentence types such as interrogatives, imperatives and exclamatives. In the texts we used to count the PoS trigram frequencies, such sentence types do not occur, which can be seen as one reason why cross-variety differences in PoS trigram frequencies are small. In sentences with a double object, the preferred (most frequent) order is recipient–object, although the positions can be optionally exchanged in most varieties. In the Guangzhou and Changsha varieties, only the reversed order is possible, i.e., object–recipient. Such structural exchanges, however, will not be picked up by our PoS trigram count, since both the recipient and the direct object are composed of the same PoS categories.

4.3. Separating Lexical and Phonetic Distance by Automatic Cognate Recognition

As discussed in Section 2.2.1, lexico-phonetic distances encompass both lexical and phonetic variation. These distances can serve as an approximate measure of mutual intelligibility between speakers of different languages, with lexical differences likely posing the greater obstacle, and phonetic variation playing a smaller role.

Nonetheless, it may be informative to decompose lexico-phonetic distances into separate measures of lexical and phonetic distance in order to more precisely assess what lexico-phonetic similarity actually reflects—that is, to estimate the relative contributions of lexical and phonetic variation. This requires a method for grouping items in the data set into cognate sets. Once such groupings are established, the lexical distance between two varieties can be calculated as the proportion of non-cognate pairs (Séguy, 1973), while phonetic distance can be defined as the mean Levenshtein distance computed over the cognate pairs alone.

An excellent tool that can do this is the LexStat function and its SCA (Sound-Class-Based Alignment) cluster method in the LingPy package. LingPy is a Python (version 2.6.13) library designed for computational historical linguistics by Johann-Mattis List (List, 2012, 2014).[9] LexStat begins by grouping input sequences into broader "sound classes" based on shared phonological features and calculating their sonority profiles. Sonority profiles help model sound change tendencies in prosodic contexts. To create language-specific scoring schemes, LexStat uses a permutation method that compares the observed distribution of aligned sound segments (from semantically similar word pairs) to an expected distribution (from semantically unrelated, randomly shuffled pairs). These distributions inform a custom scoring matrix for each language. LexStat then applies a variant of the pairwise alignment algorithm which computes the distances between all word pairs and avoids penalizing gaps at the beginning and end of sequences. Finally, words within the same semantic category are clustered into cognate sets using a modified, flat variant of the UPGMA algorithm, which halts clustering once a predefined threshold of average

pairwise distances is met. This method balances phonetic similarity with language-specific patterns to identify likely cognates.

After identifying cognate sets using the LexStat function, we computed lexical distances using Séguy's method, and phonetic distances using PMI-weighted Levenshtein distance. Subsequently, we calculated the correlation coefficients among the lexical, lexico-phonetic, and phonetic distance measures. These correlations are presented in Table 6. For completeness, we also included correlations with the syntactic distance measures. The statistical significance of all correlations was evaluated using the Mantel test.

Table 6. Correlation coefficients among lexical, lexico-phonetic, phonetic and syntactic distances. All correlations are significant ($p < 0.001$).

	Lexical	**Lexical–Phonetic**	**Phonetic**	**Syntactic**
lexical	1	0.960	0.819	0.864
lexical–phonetic	0.960	1	0.850	0.826
phonetic	0.819	0.850	1	0.589
syntactic	0.860	0.830	0.590	1

The lexico-phonetic distances show a strong correlation with the lexical distances ($r = 0.9604$), and a moderately strong correlation with the phonetic distances ($r = 0.818$). This implies that 92.2% of the variance in the lexico-phonetic distances can be attributed to lexical differences, while 67% can be attributed to phonetic differences. The correlation with lexical distances is significantly higher than that with phonetic distances ($p < 0.001$), indicating that lexico-phonetic distances primarily capture lexical rather than phonetic variation. This aligns with our earlier claim that lexical differences are likely to hinder mutual intelligibility more than phonetic differences.

The moderately strong correlation between lexical and phonetic distances suggests that these two measures capture partially distinct patterns. The dendrogram based on lexical distances closely resembles the one derived from lexico-phonetic distances, both showing the same primary division into four major groups, with only minor differences within the groups. In contrast, the dendrogram based on phonetic distances deviates more from the lexico-phonetic one. While the Germanic, Romance, Slavic and Chinese groups remain identifiable, English, Danish and French appear as outliers, diverging from the broader European cluster. Within the Chinese group, the Fuzhou, Chaozhou and Xiamen varieties cluster more closely together.

5. Conclusions

In this article, we aimed to test some specific claims made in the literature on relative distances among European languages and among Chinese dialects, suggesting that some language varieties within the Sinitic family traditionally called dialects are, in fact, more linguistically distant from one another than some European varieties that are traditionally called languages. More generally, we wished to examine whether distances among varieties within and across European language families are larger than those within and across Sinitic language varieties. To this end, we computed lexico-phonetic as well as syntactic distance measures for comparable language materials in six Germanic, five Romance and six Slavic languages, as well as for six Mandarin and nine non-Mandarin ('southern') Chinese varieties. Lexico-phonetic distances were expressed as the length-normalized MPI-weighted Levenshtein distances computed on the 100 most frequently used nouns in the 32 language varieties. Syntactic distance was implemented as the (complement of) the Pear-

son correlation coefficient found for the PoS trigram frequencies established for a parallel corpus of the same four texts translated into each of the 32 languages.

The results can be summarized as follows: both lexico-phonetic and syntactic distances are larger across language families than within families, thereby permitting the perfect classification of language varieties into their respective families. Moreover, the distances within families tend to reflect the traditional cladistic genealogy of the languages as proposed in the linguistic literature.

In general terms, we found that the lexico-phonetic distances are relatively large and of approximately equal magnitude in the Germanic, Slavic and non-Mandarin Chinese language varieties we examined. However, the lexico-phonetic distances among the Romance and Mandarin languages are considerably smaller, but also of similar magnitude.

More specifically, our results (Table 5) bear out that Cantonese (Guangzhou dialect) is lexico-phonetically as distant from Standard Mandarin (Beijing dialect) as related European languages such as Portuguese–Italian, Portuguese–Romanian and Dutch–German are. Syntactically, however, the differences among the European languages, both within and across the families, are about ten times larger than the differences within and across the groups of Sinitic varieties.

The overall conclusion of the present study is that the Sinitic language varieties—often referred to as dialects—are as different from one another as European languages are, but only in terms of lexico-phonetic distance. In their syntactic surface structure, however, the Chinese varieties are much more similar than the European languages, as shown by the exceptionally high correlations between the frequencies of PoS trigrams in the parallel text corpus. In terms of syntactic properties, then, there is every reason to consider the 15 Chinese language varieties very closely related, so the tradition to refer to them as dialects is not without justification.

This study should be regarded as a first exploration of the possibilities of quantifying linguistic distance among a wide variety of languages along multiple linguistic dimensions. We would argue that inherent cross-lingual intelligibility (i.e., in abstraction from non-linguistic factors such as prior exposure and orthographic similarity) ultimately depends on structural differences between the non-native language and the listener's own language. The differences can be organized into four linguistic domains (or dimensions), i.e., lexical, phonetic, morphological and syntactic. In the present study, we collapsed the lexical and phonological domains into a single dimension, i.e., Kessler's all-word distance. This is computationally convenient, but it fails to disentangle the separate contributions of lexical overlap and phonological transparency of the cognate word pairs. Moreover, in the present study, we analyzed segmental differences only and did not consider differences in word prosody (such as differences in stress position and in lexical tone). Differences in morphology were not included.

The scope of the materials was quite limited. Lexico-phonetic distance was computed on a list of 100 (highly frequent) nouns, while the syntactic distance was based on a (rather coarse) labeling of Part of Speech. Translators were instructed not to come up with best possible equivalent of the source text but to stick to the word order in the source text as much as possible, and to depart from the original word order only if the literal translation would violate the syntactic constraints of the target language. This choice may have led to an underestimation of the syntactic distance for languages with a relatively free or permissive word order—such as Slavic and Sinitic languages.

In spite of these shortcomings in the study, the results we obtained seem robust, and are compatible with earlier findings. The results are a reminder that the lexical overlap and transparency of cognates should not be taken as the sole criterion to argue about linguistic distance.

Author Contributions: Conceptualization, V.J.v.H.; methodology, V.J.v.H. and W.H.; software, W.H.; validation, C.T., V.J.v.H., W.H. and C.G.; formal analysis, W.H. and V.J.v.H.; resources, C.T. and C.G.; data curation, C.G., W.H. and V.J.v.H.; writing—original draft preparation, V.J.v.H. and W.H.; writing—review and editing, V.J.v.H., W.H. and C.G.; visualization, W.H. and V.J.v.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded through the editors of the special *Languages* issue on Dialectal Dynamics.

Informed Consent Statement: Informed consent was obtained from all translators involved in this study.

Data Availability Statement: The data we compiled for the present study can be downloaded from the website of the Open Science Foundation (OSF). Use the following permanent link: https://osf.io/znhyd/?view_only=85abb174a8704b31a8079bd12b9eaaf1 (accessed on 21 May 2025) and then select the files tab. The materials comprise (1) Excel sheets listing the 100 items in IPA transcription for the 32 languages in our sample, (2) Excel sheets listing the ~800-word texts with their Part of Speech tags for the 32 languages, (3) distance matrices based on lexico-phonetic similarity, (4) distance matrices based on syntactic similarity (trigram frequencies), and (5) the R-scripts used to compute the distances from the materials listed as well as the correlation coefficients between lexico-phonetic and syntactic distances. The file readme.txt provides detailed instructions on how the various plots in this article can be produced from the materials stored.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The 100 Most Frequent Nouns in the BNC Corpus and Their Equivalents in Mandarin Chinese[10]

English	Chinese	English	Chinese	English	Chinese	English	Chinese
time	时间	place	地方	development	发展	hour	小时
year	年	point	点	room	房间	rate	比率
people	人 (们)	house	房子	water	水	law	法律
man	男 (人)	country	国家	form	表格	door	门
day	(一) 天	week	周	car	汽车	court	法庭
thing	事情	member	会员	level	水平	office	办公室
child	孩子	end	尽头	policy	政策	war	战争
government	政府	word	单词	council	理事会	reason	原因
part	部分 (零件)	example	例子	line	线 (条)	minister	部长
life	生活	family	家	need	需要	subject	主题
case	情况	fact	事实	effect	效果	person	人
woman	女人	percent	百分比	use	使用	period	期间
work	工作	month	月 (份)	idea	主意	society	社会
system	系统	side	(旁) 边	study	研究	process	过程
group	组	night	夜晚	girl	女孩	mother	母亲
number	数字	eye	眼睛	name	名字	voice	声音
world	世界	head	头	result	结果	police	警察
area	区域	information	信息	body	身体	kind	种类
course	课程	question	问题	friend	朋友	price	价格
company	公司	power	权力	right	权利	position	位置
problem	问题	money	(金) 钱	authority	权威	age	年龄
service	服务	change	变化	view	视野	figure	体型
hand	手	interest	兴趣	report	报告	education	教育
party	派对	order	订单	face	脸	programme	日程
school	学校	book	书	market	市场	minute	分钟

Appendix B. The Four Short Texts in English and in Mandarin That Were Used to Compute PoS Trigram Frequencies for 32 Language Varieties[11]

Child athletes

Parents whose children show a special interest in a particular sport have a difficult decision to make. Should they allow their children to train to become top sportsmen and women? For many children it means starting very young. School work, going out with friends and other interests have to take second place. It's very difficult to explain to young children why they have to train for five hours a day. That includes even the weekend, when most of their friends are playing. Another problem is of course money. In many countries money for training is available from the government for the very best young athletes. If this help cannot be given the parents have to find the time and money to support their children. Sports clothes, transport to competitions, special equipment etc. can all be very expensive. Many parents are understandably worried that it's dangerous to start serious training in a sport at an early age. Some doctors agree that young muscles may be damaged by training before they are properly developed. Trainers, however, believe that you can only reach the top as a sports person when you start young. What is clear is that very few people do reach the top. So both parents and children should be prepared for failure. It happens even after many years of training.

对于 那些 孩子 对 某项 特定 运动 表现 出 特别 兴趣 的 父母 来说，他们 需要 做出 一个 艰难 的 决定 。他们 是否 应该 允许 孩子 接受 训练， 成为 顶级 运动员？ 对于 许多 孩子 来说，这 意味 着 从 很 小 的 时候 就 开始 学业、与 朋友 出去 玩 和 其他 兴趣 都 必须 放 在 第二位。很 难 向 年幼 的 孩子 解释 为什么 他们 每天 要 训练 五个 小时， 甚至 包括 周末，这 时候 他们 的 大部分 朋友 都 在 玩耍。另一 个 问题 当然 是 金钱。在 许多 国家， 政府 为 最 优秀 的 年轻 运动员 提供 训练 资金。如果 得不到 这种 资助， 父母 就 必须 找到 时间 和 金钱 来 支持 孩子。运动 服装、比赛 的 交通费、特殊 装备 等 都 可能 非常 昂贵。许多 父母 担心 在 孩子 年幼 时 进行 严格 的 运动 训练 可能 会 对 肌肉 造成 损伤。一些 医生 同意， 在 肌肉 适当 发育 之前 进行 训练 可能 会 对 年幼 的 肌肉 造成 损害。然而， 教练们 相信， 只有 在 年幼 时 训练 才能 成为 顶级 运动员。显而易见 的 是， 很少 有人 能 达到 顶峰。 因此， 父母 和 孩子 都 应该 为 失败 做 好 准备。即使 经过 多年 的 训练， 失败 也 是 有 可能 发生 的。

Catching a cold

Hello, my name is Bill and I give advice to people with questions about their health. I get a lot of letters at this time of year. People complain that they have a cold which won't go away. There are so many different stories about how to prevent or cure a cold. So it's often difficult to know what to do. Colds are rarely dangerous, except for people who are already weak, such as the elderly or young babies. Still, colds are always uncomfortable and usually most unpleasant. Of course you can buy lots of medicines which will help to make your cold less unpleasant. But remember that nothing can actually cure a cold or make it go away faster. Another thing is that any medicine which is strong enough to make you feel better could be dangerous. If you are already taking medicine for other illnesses always check with your doctor if that's all right. And remember that it could happen that they might make you sleepy. Please don't try to drive if they do! Lastly, there are no magic foods or drinks. The best answer is to keep strong and healthy. You'll have less chance of catching a cold, and if you do, it shouldn't be so bad!

你好， 我的 名字 叫 克里斯蒂娜， 我 为 有 健康 疑问 的 人 提供 建议。每年 此 时， 我 都 会 收到 许多 来信。人们 抱怨 他们 的 感冒 不 消退。关于 预防 或 治疗 感冒， 有 很多 不同 的 说法， 所以 往往 很 难 知道 该 怎么 做。感冒 很少 有 危险， 除非 对 本就 虚弱 的 人，如 老人 或 婴儿。 然而， 感冒 总是 让 人 不 舒服， 而且 通常 也 很 不 愉快。当然，你 可以 买 很多 药， 这 将 有助于 减轻 感冒 的 不适感。但是 请 记住， 没有 药物 能 真正 治愈 感冒， 或者 让 它 好 得 更 快。另外， 任何 能 让 你 感觉 好些 的 药物 都 可能 是 危险的。如果 你 已经 在 服用 治疗 其他 疾病 的 药物， 请 务必 咨询 你 的 医生， 与 感冒药 一同 服用 是否 可以。而且 请 记住， 药物 可能 会 让 你 昏昏欲睡。如果 出现 这种 情况，

请 不要 尝试 开车！ 最后， 没有 什么 神奇 的 食物 或 饮料。最好 的 办法 就是 保持 强壮 和 健康。 这样 你 感冒 的 机会 会 更 小， 即使 感冒 了， 也 不会 那么 糟糕！

Driving in Winter

Winter is dangerous because it's so difficult to know what is going to happen. Accidents take place so easily. Fog can be waiting to meet you over the top of a hill. Ice might be hiding beneath the melting snow, waiting to send you off the road. The car coming towards you may suddenly slide across the road. Rule Number One for driving on icy roads is to drive smoothly. Uneven movements can make a car suddenly very difficult to control. Every time you turn the wheel, brake or increase speed, you must be gentle and slow as possible. Imagine you are driving with a full cup of hot coffee on the seat next to you. Drive so that you wouldn't spill it. Rule Number Two is to pay attention to what might happen. The more ice there is, the further down the road you have to look. Test how long it takes to stop by gently braking. Remember that you may be driving more quickly than you think. In general, allow double your normal stopping distance when the road is wet. Use three times this distance on snow, and even more on ice. Try to stay in control of your car at all times and you will avoid trouble.

冬季 (驾驶) 很 危险， 因为 很 难 知道 会 发生 什么 事情。 事故 很 容易 发生。 雾 可能 在 山顶 等 着 见 你。冰 可能 隐藏 在 融化 的 雪 下， 等 着 把 你 甩 出 路面。向 你 驶 来 的 车 可能 会 突然 滑 过 道路。在 冰雪 路面 驾驶 的 第一 规则 是 平稳 驾驶。不 平稳 的 动作 会 使 车辆 突然 难以 控制。每次 你 转动 方向盘、 刹车 或 加速 时， 都 要 尽量 轻柔 和 缓慢。想象 一下， 你 正在 开车，满满 一杯 热 咖啡 放 在 旁边 的 座位 上。平稳 地 驾驶， 确保 你 别 撒 了 咖啡。第二 规则 是 要 注意 可能 发生 的 情况。冰面 越 多， 你 就 越 需要 更 远 地 向前 看。 通过 轻轻 刹车 来 测试 停车 需要 多长 时间。记住， 你 驾驶 的 速度 可能 比 你 想象 的 要 快。一般来说， 当 路面 湿滑 时， 允许 你 的 停车 距离 是 平时 的 两倍。在 雪地 上 则要 增加 三倍 的 停车 距离， 在 冰面 上 要 更 多 (距离)。尽量 一直 控制 好 你 的 车， 这样 你 就 能 避免 麻烦。

Riding a bike

Getting enough exercise is part of a healthy lifestyle. Along with jogging and swimming, riding a bike is one of the best all-round forms of exercise. It can help to increase your strength and energy. Also it gives you more efficient muscles and a stronger heart. But increasing your strength is not the only advantage of riding a bike. You're not carrying the weight of your body on your feet. That's why riding a bike is a good form of exercise for people with painful feet or backs. However, with all forms of exercise it's important to start slowly and build up gently. Doing too much too quickly can damage muscles that aren't used to working. If you have any doubts about taking up riding a bike for health reasons, talk to your doctor. Ideally you should be riding a bike at least two or three times a week. For the exercise to be doing you good, you should get a little out of breath. Don't worry that if you begin to lose your breath, it could be dangerous. This is simply not true. Shortness of breath shows that the exercise is having the right effect. However, if you find you are in pain then you should stop and take a rest. After a while it will get easier.

保持 充足 的 运动 是 健康 生活 方式 的 一 部分。 除了 慢跑 和 游泳， 骑 自行车 也 是 最好 的 全方位 锻炼 方式 之一。它 可以 帮助 增加 你 的 力量 和 能量， 使 你 的 肌肉 更加 高效， 心脏 更 强壮。但 增加 力量 并 不是 骑 自行车 的 唯一 好处。 你 的 脚 没有 承受 你 身体 的 重量。这 就是 为什么 骑 自行车 对于 脚痛 或 背痛 的 人 来说 是 一种 很好 的 锻炼 方式。然而， 对于 所有 形式 的 锻炼 来说， 缓慢 开始、 逐渐 增加 是 很 重要的。过快 过多 的 运动 可能 会 损伤 不习惯 工作 的 肌肉。 如果 你 因 健康 原因 对 骑 自行车 有 任何 疑虑， 请 咨询 医生。理想 情况 下， 你 应该 每周 至少 骑 自行车 两 到 三次。为了 使 锻炼 对 你 有益， 你 应该 稍微 有些 气喘。 不要 担心 如果 你 开始 喘 不过 气， 这 可能 是 危险的。这 完全 不是 真的。呼吸 急促 表明 锻炼 产生 了 正确 的 效果。然而， 如果 你 发现 自己 感到 疼痛， 那 就 应该 停 下来 休息 一下。过 一段 时间 后， 情况 会 变 得 更 容易。

Appendix C

Upper triangle: Lexico-phonetic distance matrix based on PMI-weighted Levenshtein distance for 17 European languages.

Lower triangle: Syntactic distance matrix based on PoS trigram frequencies.

	Bu	Cr	Cz	Da	Du	En	Fh	Fn	Ge	It	Po	Pt	Ro	Sk	Sn	Sp	Sw
Bu		0.377	0.515	0.592	0.563	0.588	0.597	0.577	0.594	0.548	0.512	0.510	0.514	0.509	0.475	0.547	0.591
Cr	0.330		0.464	0.616	0.584	0.632	0.613	0.587	0.580	0.561	0.490	0.586	0.554	0.470	0.368	0.551	0.605
Cz	0.317	0.217		0.627	0.605	0.614	0.603	0.598	0.591	0.576	0.385	0.614	0.591	0.209	0.451	0.603	0.595
Da	0.417	0.396	0.378		0.512	0.583	0.605	0.526	0.518	0.602	0.628	0.594	0.577	0.628	0.614	0.588	0.475
Du	0.547	0.482	0.458	0.257		0.543	0.570	0.289	0.427	0.554	0.602	0.572	0.538	0.609	0.583	0.580	0.471
En	0.440	0.422	0.389	0.188	0.269		0.580	0.546	0.570	0.569	0.613	0.558	0.545	0.626	0.618	0.580	0.546
Fh	0.664	0.567	0.558	0.338	0.354	0.261		0.571	0.610	0.490	0.591	0.471	0.508	0.600	0.616	0.508	0.581
Fn	0.558	0.491	0.465	0.275	0.042	0.274	0.363		0.474	0.560	0.601	0.579	0.541	0.616	0.588	0.582	0.478
Ge	0.569	0.489	0.456	0.295	0.162	0.302	0.376	0.185		0.580	0.588	0.601	0.555	0.589	0.594	0.599	0.484
It	0.587	0.496	0.487	0.377	0.381	0.314	0.213	0.372	0.405		0.570	0.449	0.436	0.588	0.581	0.345	0.566
Po	0.314	0.215	0.198	0.354	0.449	0.393	0.563	0.468	0.446	0.483		0.604	0.581	0.386	0.502	0.605	0.593
Pt	0.543	0.454	0.431	0.331	0.366	0.286	0.229	0.368	0.396	0.174	0.445		0.479	0.616	0.602	0.390	0.577
Ro	0.504	0.396	0.391	0.314	0.380	0.279	0.252	0.384	0.404	0.228	0.370	0.199		0.602	0.575	0.458	0.569
Sk	0.309	0.202	0.095	0.359	0.455	0.395	0.557	0.464	0.469	0.487	0.185	0.443	0.383		0.463	0.608	0.603
Sn	0.286	0.217	0.252	0.376	0.480	0.403	0.569	0.502	0.469	0.530	0.232	0.483	0.444	0.261		0.578	0.598
Sp	0.584	0.505	0.501	0.347	0.377	0.291	0.185	0.372	0.404	0.148	0.497	0.151	0.209	0.494	0.520		0.582
Sw	0.406	0.367	0.359	0.152	0.277	0.198	0.339	0.296	0.310	0.349	0.333	0.337	0.315	0.356	0.366	0.348	

Bu: Bulgarian, Cr: Croatian, Cz: Czech, Da: Danish, Du: Dutch, En: English, Fh: French, Fn: Frisian, Ge: German, It: Italian: Po: Polish, Pt: Portuguese: Ro: Romanian, Sk: Slovak, Sn: Slovenian, Sp: Spanish, Sw: Swedish.

Appendix D

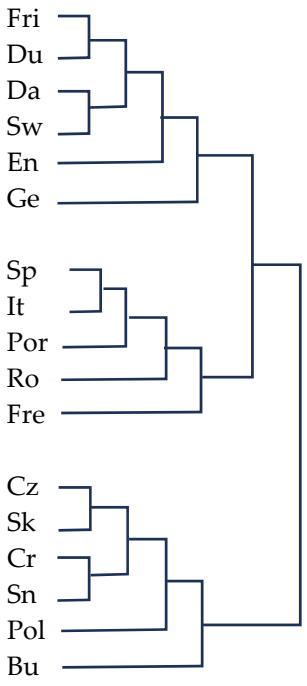
Upper triangle: Lexico-phonetic distance matrix based on PMI-weighted Levenshtein distance for 15 Chinese dialects

Lower Triangle: Syntactic distance matrix based on PoS trigram frequencies.

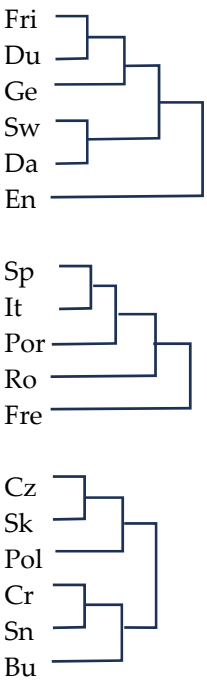
	Be	Cs	Cz	Cd	Fu	Gu	Ji	Me	Na	Su	Ta	We	Wu	X'n	Xm
Beijing		0.335	0.485	0.288	0.558	0.473	0.333	0.482	0.411	0.513	0.346	0.514	0.344	0.329	0.551
Changsha	0.032		0.459	0.230	0.544	0.438	0.330	0.443	0.369	0.433	0.309	0.461	0.291	0.312	0.548
Chaozhou	0.031	0.001		0.447	0.465	0.449	0.503	0.454	0.461	0.525	0.425	0.487	0.483	0.483	0.466
Chengdu	0.015	0.036	0.035		0.525	0.432	0.326	0.405	0.328	0.438	0.255	0.462	0.191	0.301	0.513
Fuzhou	0.021	0.038	0.037	0.023		0.487	0.556	0.478	0.505	0.580	0.503	0.549	0.535	0.552	0.528
Guangzhou	0.009	0.030	0.029	0.011	0.015		0.499	0.424	0.414	0.518	0.457	0.478	0.430	0.487	0.505
Jinan	0.010	0.031	0.030	0.012	0.019	0.007		0.496	0.431	0.502	0.305	0.480	0.389	0.199	0.554
Meixian	0.008	0.025	0.024	0.011	0.014	0.005	0.005		0.380	0.514	0.453	0.508	0.430	0.485	0.482
Nanchang	0.008	0.025	0.024	0.011	0.014	0.005	0.005	0.000		0.468	0.403	0.477	0.345	0.420	0.507
Suzhou	0.011	0.028	0.027	0.014	0.017	0.008	0.002	0.004	0.004		0.448	0.459	0.438	0.517	0.571
Taiyuan	0.007	0.025	0.024	0.011	0.014	0.005	0.006	0.000	0.000	0.003		0.439	0.351	0.263	0.548
Wenzhou	0.007	0.025	0.024	0.011	0.014	0.005	0.006	0.000	0.000	0.003	0.000		0.495	0.474	0.578
Wuhan	0.012	0.033	0.032	0.013	0.021	0.007	0.003	0.008	0.008	0.004	0.008	0.008		0.374	0.524
Xi'an	0.007	0.025	0.024	0.011	0.014	0.005	0.006	0.000	0.000	0.003	0.000	0.000	0.008		0.567
Xiamen	0.009	0.021	0.020	0.012	0.015	0.006	0.007	0.002	0.002	0.005	0.002	0.002	0.009	0.002	

Appendix E

Comparison of affinity trees based on lexico-phonetic distance. The trees in the left-hand column are based on Müller et al. (2021), for European languages only. Those in the right-hand column are based on our own study (Figure 1, European languages only). The trees below preserve the branching found in the source figures, but the levels of attachment have been abstracted to facilitate visual comparison. For abbreviations, see Appendix C (note).



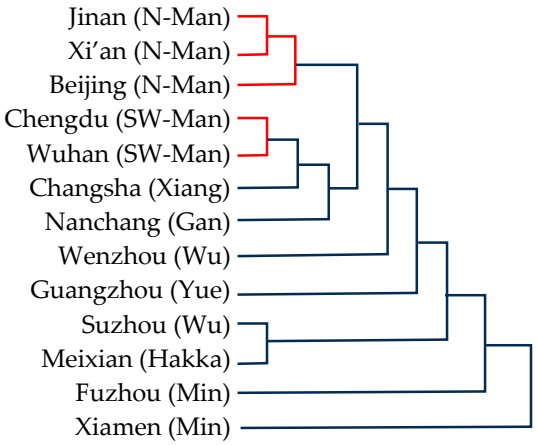
Polyakov et al. (2009)



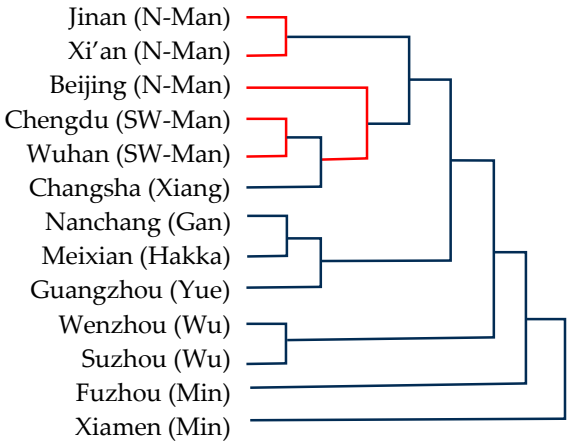
Present study

Appendix F

Side-by-side comparison of the affinity trees extracted from Müller et al. (2021) based on lexico-phonetic distances computed for Sinitic varieties and our own data (Figure 1, Sinitic varieties). The Mandarin varieties are in red print. The trees preserve the branching found in the source figures, but the levels of attachment have been abstracted to facilitate visual comparison.



Müller et al. (2021)



Present study

Notes

1 Appendices E and F contain a comparison of the affinity trees produced on the basis of the ASJP database and ours. The lexico-phonetic distances computed by both methods for the 32 language varieties at issue correlate at $r = 0.925$. See also Section 4.

2 This claim may not be true in the case of onomatopoeic words and word forms based on sound symbolism. As a precaution, Wichmann et al. (2010) therefore propose that the mean length-normalized Levenstein distance (LDN) computed on all cognate pairs be divided by the mean LDN for all non-cognate pairs. This ratio (LDND) was found to be a slightly (but not significantly) better predictor of cladistic language trees than LDN.

[3] West Frisian is a West Germanic language spoken as the native language by an estimated 415,000 citizens in the north of the Netherlands (Robinson-Jones & Scarse, 2022, p. 3). It has an official status along with and equal to Dutch in the province of Friesland/Fryslân. The language was included in the present study by special request of the Frisian Academy.

[4] We used LED-A for calculating the PMI Levenshtein distances. The implementation of PMI–Levenshtein in LED-A differs in its details from Wieling (2012). See https://www.led-a.org/docs/PMI.pdf (accessed on 21 May 2025).

[5] See also: https://www.led-a.org/docs/Diacr.pdf (accessed on 21 May 2025).

[6] See also: https://www.led-a.org/docs/Ngram.pdf (accessed on 21 May 2025).

[7] This result mirrors earlier findings by Gooskens and Heeringa (2004), computed on phonetically transcribed readings of the Aesop *North wind and sun* fable, which clearly showed that, contrary to popular belief, Frisian is (much) closer to Dutch than to English—or any other Germanic language. It is true, nevertheless, that the distance between English and Frisian is (slightly) smaller than between English and Dutch—both in our and earlier results. The urban legend is perpetuated by selective sampling of the striking similarities between English and Frisian, such as those mentioned by Pei (1966, p. 34).

[8] We also checked this using the Mantel test and obtained the same findings.

[9] See https://lingpy.org/ (accessed on 21 May 2025).

[10] The parallel lists of the orthographic form, IPA transcription and (for Chinese) Romanized Pinyin (including lexical tone) of the 100 items for each of the 32 language varieties can be downloaded from the OSF site (see section "Data Availability Statement")

[11] The complete texts with PoS tags added after each word can be downloaded from the OSF site (see section "Data Availability Statement").

References

BNC Consortium. (2007). *BNC: The British National Corpus, version 3, BNC XML edition*. Oxford University Computing Services on Behalf of the BNC Consortium. Available online: http://www.natcorp.ox.ac.uk/ (accessed on 21 May 2025).

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Liber. Available online: https://portal.research.lu.se/en/publications/311242e6-5f2b-489a-b35d-5ffef5ec5301 (accessed on 21 May 2025).

Ceolin, A., Guardiano, C., Irimia, M. A., & Longobardi, G. (2020). Formal syntax and deep history. *Frontiers in Psychology*, *11*, 488871. [CrossRef] [PubMed]

Ceolin, A., Guardiano, C., Longobardi, G., Irimia, M. A., Bortolussi, L., & Sgarro, A. (2021). At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society*, *376*, 20200197. [CrossRef] [PubMed]

Cheng, C.-C. (1997). Measuring relationship among dialects: DOC and related resources. *Computational Linguistics & Chinese Language Processing*, *2*(1), 41–72. Available online: https://aclanthology.org/O97-3002/ (accessed on 21 May 2025).

Council of Europe. (2001). *Common European framework of reference for languages. Learning, teaching, assessment*. Cambridge University Press. Available online: https://rm.coe.int/1680459f97 (accessed on 21 May 2025).

Di Buccio, E., Di Nunzio, G. M., & Silvello, G. (2014). A vector space model for syntactic distances between dialects. In *Language resources and evaluation conference* (pp. 2486–2489). ELRA. Available online: https://aclanthology.org/L14-1148/ (accessed on 21 May 2025).

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online (v2020.4) [data set]*. Zenodo. Available online: https://wals.info (accessed on 21 May 2025). [CrossRef]

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*, 79–82. [CrossRef]

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, *309*, 2072–2075. [CrossRef]

Dyen, I., Kruskal, J. B., & Black, P. (1992). *An Indoeuropean classification: A lexicostatistical experiment* (Transactions of the American Philosophical Society 82.5). University of Pennsylvania Press. Available online: http://www.ldc.upenn.edu/ (accessed on 21 May 2025). [CrossRef]

Gårding, E. (1974). *Kontrastiv prosodi*. Gleerup.

Golubović, J. (2016). *Mutual intelligibility in the Slavic language area*. Center for Language and Cognition. Available online: https://research.rug.nl/files/31880596/Complete_thesis.pdf (accessed on 21 May 2025).

Gooskens, C. (2024). *Mutual intelligibility between closely related languages*. De Gruyter Mouton. [CrossRef]

Gooskens, C., & Heeringa, W. (2004). The position of Frisian in the Germanic language area. In D. G. Gilbers, M. J. Schreuder, & N. Knevel (Eds.), *On the boundaries of phonology and phonetics* (pp. 61–87). Department of Linguistics, Groningen University. Available online: https://www.researchgate.net/publication/237534065 (accessed on 21 May 2025).

Gooskens, C., & van Heuven, V. J. (2017). Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups. *Speech Communication*, *89*, 25–36. [CrossRef]

Gooskens, C., & van Heuven, V. J. (2020). How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? *Linguistic Approaches to Bilingualism*, *10*(3), 351–379. [CrossRef]

Gooskens, C., & van Heuven, V. J. (2021). Mutual intelligibility. In M. Zampieri, & P. Nakov (Eds.), *Similar languages, varieties, and dialects: A computational perspective* (pp. 51–95). Cambridge University Press. [CrossRef]

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., & Voigt, S. (2018). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, *15*(2), 169–193. [CrossRef]

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*, 435–439. [CrossRef]

Gray, R. D., Bryant, D., & Greenhill, S. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society*, *365*, 3923–3933. [CrossRef]

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (Groningen Dissertations in Linguistics 46). Groningen Centre for Language and Cognition. Available online: https://research.rug.nl/files/9800656/thesis.pdf (accessed on 21 May 2025).

Heeringa, W., Gooskens, C., & van Heuven, V. J. (2023). Comparing Germanic, Romance and Slavic: Relationships among linguistic distances. *Lingua*, *287*, 1–23. [CrossRef]

Heeringa, W., Swarte, F., Schüppert, A., & Gooskens, C. (2018). Measuring syntactical variation in Germanic texts. *Digital Scholarship in the Humanities*, *33*(1), 279–296. [CrossRef]

Inkelas, S., & Zec, D. (1988). Serbo-croatian pitch accent: The interaction of tone, stress and intonation. *Language*, *64*(2), 227–248. [CrossRef]

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall.

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, *3*(2), 245–291. [CrossRef]

Kessler, B. (1995, March 27–31). *Computational dialectology in Irish Gaelic*. 7th Conference of the European Chapter of the Association for Computational Linguistics (pp. 60–67), Dublin, Ireland. Available online: https://aclanthology.org/E95-1009/ (accessed on 21 May 2025).

Li, C., & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, R. (Ed.). (1993–1999). *The comprehensive dictionaries of modern Chinese dialects*. Jiang Su Education Publishing House.

List, J.-M. (2012, April 23–24). *LexStat: Automatic detection of cognates in multilingual wordlists*. EACL 2012 Joint Workshop of LINGVIS & UNCLH (pp. 117–125), Avignon, France. Available online: https://aclanthology.org/W12-0216.pdf (accessed on 21 May 2025).

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf University Press. [CrossRef]

Longobardi, G. (2005). A minimalist program for parametric linguistics. In H. Broekhuis, N. Corver, R. Huijbregts, U. Kleinhenz, & J. Koster (Eds.), *Organizing grammar* (pp. 407–414). DeGruyter Mouton. Available online: https://people.sissa.it/~ale/EU_infoday/Lon05.pdf (accessed on 21 May 2025).

Mian Yan, M. (2006). *Introduction to Chinese dialectology* (LINCOM Studies in Asian Linguistics). LINCOM. Available online: https://starlingdb.org/Texts/Students/Mian%20Yan,%20Margaret/Introduction%20to%20Chinese%20Dialectology%20(2006).pdf (accessed on 21 May 2025).

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*(3), 161–175. [CrossRef]

Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Holman, E. W., Sauppe, S., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., … Valenzuela, P. (2021). *ASJP world language trees of lexical similarity: Version 5 (October 2021)*. Available online: https://asjp.clld.org/static/WorldLanguageTree-005.zip (accessed on 21 May 2025).

Nerbonne, J., & Wiersma, W. (2006). A measure of aggregate syntactic distance. In J. Nerbonne, & E. Hinrichs (Eds.), *Proceedings of the workshop on linguistic distances* (pp. 82–90). Association for Computational Linguistics. Available online: https://aclanthology.org/W06-1111.pdf (accessed on 21 May 2025).

Pei, M. (1966). *The story of language*. Allen & Unwin.

Polyakov, V. N., Solovyev, V. D., Wichmann, S., & Belyaev, O. (2009). Using WALS and Jazyki Mira. *Linguistic Typology*, *13*, 135–165. [CrossRef]

Pompei, S., Loreto, V., & Tria, F. (2011). On the accuracy of language trees. *PLoS ONE*, *6*(6), e20109. [CrossRef]

Riad, T. (2014). *The phonology of Swedish*. Oxford University Press.

Robinson-Jones, C., & Scarse, Y. R. (2022). *Report on the West Frisian language (Language technology support of Europe's languages in 2020/2021—European language equality project)*. Available online: https://www.researchgate.net/publication/361644854 (accessed on 21 May 2025).

Séguy, J. (1973). La dialectometrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane*, *37*, 1–24.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*, 33–40. [CrossRef]

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*, 452–463. Available online: https://www.jstor.org/stable/i357803 (accessed on 21 May 2025).

Swarte, F. H. E. (2016). *Predicting the (mutual) intelligibility of Germanic languages from linguistic and extralinguistic factors*. Center for Language and Cognition. Available online: https://research.rug.nl/files/29253828/Complete_thesis.pdf (accessed on 21 May 2025).

Tang, C. (2009). *Mutual intelligibility of Chinese dialects. An experimental approach* (LOT dissertation Series 228). LOT. Available online: https://www.lotpublications.nl/Documents/228_fulltext.pdf (accessed on 21 May 2025).

Tang, C., & van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, *119*, 709–732. [CrossRef]

Tang, C., & van Heuven, V. J. (2015). Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics*, *52*(3), 285–311. [CrossRef]

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

van Heuven, V. J., Gooskens, C. S., & van Bezooijen, R. (2015). Introducing MICRELA: Predicting mutual intelligibility between closely related languages in Europe. In J. Navracsics, & S. Bátyi (Eds.), *First and second language: Interdisciplinary approaches* (pp. 127–145). Tinta Könyvkiadó. Available online: https://www.let.rug.nl/gooskens/pdf/publ_almadi_2015.pdf (accessed on 21 May 2025).

Wardhaugh, R. (2008). *An introduction to sociolinguistics* (6th ed.). Blackwell.

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A*, *389*(17), 3632–3639. [CrossRef]

Wichmann, S., Holman, E. W., & Brown, C. H. (Eds.). (2022). *The ASJP database (version 20)*. Available online: http://asjp.clld.org/ (accessed on 21 May 2025).

Wichmann, S., & Saunders, A. (2007). How to use typological databases in historical linguistic research. *Diachronica*, *24*(2), 373–404. [CrossRef]

Wieling, M. (2012). *A quantitative approach to social and geographical dialect variation*. Centre for Language and Cognition. Available online: https://research.rug.nl/en/publications/cd637817-572f-4826-98c1-08272775fb64 (accessed on 21 May 2025).

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307–314. [CrossRef]

Wieling, M., Prokić, J., & Nerbonne, J. (2009). Evaluating the pairwise alignment of pronunciations. In L. Borin, & P. Lendvai (Eds.), *Proceedings of the EACL 2009 workshop on language technology and resources for cultural heritage, social sciences, humanities, and education (LaTeCH—SHELT&R 2009)* (pp. 26–34). Association for Computational Linguistics. Available online: https://aclanthology.org/W09-0304/ (accessed on 21 May 2025).

Yang, C., & Castro, A. (2008). Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing*, *2*(1–2), 205–219. [CrossRef]

Yip, M. (2002). *Tone*. Cambridge University Press.

Zhang, X. (Ed.). (2009). *New dictionary of Chaoshan dialect (bilingual dictionary of Mandarin-Chaozhou dialect)*. Guangdong People's Publishing House.

Zhang, Z. (2003). 现代汉语方言语序问题的考察 [An investigation of word order across Chinese dialects]. *Fāngyán*, *25*(2), 108–126.